# 在线社交网络文本内容对抗技术

刘晓明 张兆晗 杨晨阳 张宇辰 沈 超 周亚东 管晓宏

(西安交通大学电子与信息学部 西安 710049)

**摘 要** 在线社交网络内容对抗技术是人工智能领域与网络空间安全领域的一个新兴研究方向.它是指人们基于特定任务,在社交网络受众广泛、内容数量庞大、内容质量参差、内容真伪难辨的环境下,利用新兴的大数据驱动的人工智能方法,自动完成在线社交网络中针对特定主题与群体的对抗内容发现、生成与投送,进而实现社交平台异常信息的检测与反制,以达到维护网络空间安全的目的.虽然在线社交网络对抗技术属于一个新概念,鲜有相关工作,但是已有的机器学习方法可以被应用到该领域,通过特征提取、文本模式解析、文本内容编码与重建、目标优化等技术对社交文本大数据进行解析与内容表示,解决网络空间中的文本内容安全问题,实现对社交网络文本环境的净化.此外,在社交网络文本内容对抗的过程中,对抗双方的策略可作为反馈信息,使对抗模型不断进行更新和优化,最终达到完善模型的目的.基于以上攻防对抗思想,本文着重从文本内容生成与检测两方面对在线社交网络对抗进行阐述.首先,本文介绍了有关在线社交网络文本对抗技术的相关基础知识.其次,针对社交网络文本内容检测方法,本文从基于零次分类器的模型、基于机器特征的模型、基于预训练语言模型的方法、基于人机协作的模型、基于能量基础的模型 5 个角度进行详细介绍.为了方便读者针对不同的应用场景选择合适的模型,本文对不同检测模型的适用场景以及模型优劣进行了对比总结.针对社交网络文本生成方法,本文对基于对抗生成网络的文本生成模型、可控文本生成模型、长文本生成、文本质量评价 4 方面进行了综述.此外,为了方便读者对模型的有效性进行验证,本文对相关数据进行了系统性总结.最后,本文总结了在线社交网络对抗技术未来的重要研究方向与挑战.

**关键词** 社交网络对抗;人工智能;网络空间安全;文本内容自动生成;机器生成内容检测

**中图法分类号** TP391 **DOI 号** 10.11897/SP.J.1016.2022.01571

# Adversarial Technology of Text Content on Online Social Networks

LIU Xiao-Ming ZHANG Zhao-Han YANG Chen-Yang ZHANG Yu-Chen
SHEN Chao ZHOU Ya-Dong GUAN Xiao-Hong

(*Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an* 710049)

**Abstract** With prosperous development of social network and intensive research on natural language processing technology, adversarial technology of text content on online social networks becomes an emerging research direction in the field of artificial intelligence and cyberspace security. Adversarial technology of text content on online social networks refers to that in social network environment where there exists numerous users, huge amount of text contents, uneven quality of texts, ambiguous credibility of information, people employ the latest artificial intelligence methods to accomplish specific tasks such as discovery of machine-generated text, coherent and controllable

收稿日期:2020-12-26;在线发布日期:2021-08-16.本课题得到国家自然科学基金(61902308,62103323,61822309,61773310,U1736205,U1766215)、博士后创新人才支持计划基金(BX20190275,BX20200270)、博士后面上基金(2019M663723,2021M692565)、西安交通大学基本科研经费(xjh032021058,xxj022019016,xtr022019002)资助.**刘晓明**,博士,副教授,中国计算机学会(CCF)会员,主要研究方向为大规模图数据挖掘、异构数据分析、社交网络对抗、机器学习方法及其应用.E-mail:xm.liu@xjtu.edu.cn.**张兆晗**,硕士研究生,主要研究方向为社交网络对抗、机器生成文本检测.**杨晨阳**,硕士研究生,主要研究方向为社交网络对抗、可控文本自动生成.**张宇辰**,硕士研究生,主要研究方向为社交网络对抗、机器生成文本检测.**沈 超**(通信作者),博士,教授,中国计算机学会(CCF)高级会员,主要研究领域为网络物理系统优化和安全性、网络和系统安全性以及人工智能安全性.E-mail:chaoshen@xjtu.edu.cn.**周亚东**,博士,副教授,主要研究方向为数据分析与挖掘、网络科学及其应用.**管晓宏**,博士,教授,中国科学院院士,IEEE Fellow,主要研究领域为复杂网络分配与调度、网络安全、传感器网络.

text generation and adversarial content delivery for specific topics and targeted groups in online social networks automatically and precisely to filter the content in online social networks based on analysis of generation strategies, appropriate representation pattern of aimed groups, which could realize the detection and countermeasures of social platform abnormal information attack, increase the information credibility in social network, improve the quality of daily news received by social network users, enhance public trust on social media and protect cyberspace from information bombing and misleading information by reliable and economic means. Although online social network adversarial technology is a new concept with few related work, existing machine learning methods can be applied to this field by applying advanced technologies such as feature extraction, document parsing, encoding and reconstruction of text content, objective optimization on large scale social network text content dataset to solve the practical problem and clear the Internet environment. Besides, during the adversarial process of text content on online social networks, the strategies of the opposing parties can be used as feedback information to each other, so that the adversarial model is continuously updated and optimized, and finally the goal of perfecting the model is achieved. Based on the adversarial idea of attack and protection, this paper mainly describes the online social network adversarial technology from text content generation and machine-generated text content detection, respectively. Firstly, this paper introduces some basic knowledge about deep learning which is related to the online social network adversarial technology, including basic deep learning networks, pre-trained models and advanced deep learning networks. For generated social network content detection methods, this paper introduces it from different perspectives in detail, including zero-shot learning based models, machine feature based models, pre-trained language model based models, human-computer collaboration based models, and energy based models. For the convenience of choosing appropriate method to practice text content detection method in different scenes for readers, this paper makes comparison among different detection models in terms of applicable scenes, advantages and disadvantages. For social network text auto-generation methods, this paper summarizes four aspects of work, including text generation models based on adversarial generation networks, controllable text generation models, long text generation, and generated text quality evaluation. Besides, to help readers put the models into practice, verify the validity of the model and improve the shortcoming and performance of model with ease, this paper systematically summarizes the relevant datasets. Finally, this paper summarizes the important research directions and challenges of online social network adversarial technology in the future.

**Keywords** social network adversarial technology; artificial intelligence; cyberspace security; auto-generated text; machine-generated content detection

## 1 引 言

在线社交网络已经成为了人们生活中不可或缺的一部分[1]. 一项调研① 显示,在 2020 年全球有超过 36 亿用户正在使用社交媒体平台,每个用户每天在社交网络与通信软件上平均花费 144 分钟. 其中,信息交流作为社交网络平台重要的功能之一,可以使得任何人只要通过一台与互联网连接的设备,就可以将信息与全球共享. 因此,无论是专业人士还是普通大众,都倾向于从社交网络中获取最新的信息,追踪紧急事件的发展,了解公众观点和关注国际形势. 但是也正因为社交网络平台的受众广泛,出现在社交网络上的不良信息会对现实世界造成真实且严重的影响. 例如,2013 年 4 月 23 日,叙利亚电子信息

① https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/

部队入侵了美联社推特官方账号，发布了关于对白宫恐怖袭击的虚假谣言，并声称奥巴马总统在恐怖袭击中身受重伤[2]. 此消息造成了美国股市道琼斯指数在几分钟之内暴跌了 1000 点(约 9%)，是历史上最大的单日跌幅. 由此可见，在线社交网络安全在我们日常生活中扮演着至关重要的角色.

随着人工智能技术的发展，尤其是随着大规模预训练语言模型 BERT[3]、GPT-2[4]、GPT-3[5] 的出现，人们已经能够以较低门槛使用机器批量生产以假乱真的文本内容. 最新的文本生成技术包括故事写作[6-8]、对话生成[9-11]、文本总结[12-14]、医学报告的生成[15-16]以及高考作文的写作[17]等. 这些研究为人们的生活带来便利和乐趣，但是别有用心的群体也会利用这项技术的低成本特性和在线社交网络的强大传播能力，大量生产虚假新闻[18-19]、虚假产品评论[20]、诈骗信息①[21]等危害网络空间安全. 据 OpenAI 报道②，目前已有超过 300 个 APP 使用了 GPT-3，GPT-3 每天平均产出 45 亿个单词. 由此可见，语言模型如今已进入了大规模应用阶段. 已有工作表明，文本内容占据社交网络媒体内容总量的约 80%③，语言模型，尤其是机器文本内容生成模型的滥用，无疑严重影响了正常用户的生活，降低了网络消息的可信度.

不仅如此，社交媒体中的虚假新闻几乎占所有新闻消费的 6%[22]. 此外，剑桥分析公司通过分析 5000 万 Facebook 用户，为美国 2016 年特朗普总统大选的胜利提供了政治建议和内容推送策略. 在此之后，土耳其、墨西哥和印度大选中都发现了机器生成内容对选民意向进行引导，进而影响大选结果的现象[23]. 因此，这些恶意操纵社交网络内容生成的行为严重危害着网络空间安全，也违背了技术提出的初衷.

基于以上网络空间安全问题与挑战，我们针对在线社交网络对抗技术，尤其是社交网络平台中文本内容的自动生成与检测技术，展开全面调研. 本文中文本内容自动生成是指针对在线社交网络中特定主题、特定群体、特定用户情感与反馈，使用机器学习模型，生成定制化文本内容. 相对地，机器生成文本内容检测是指针对在线社交网站文本内容，从语言结构、文本质量、生成策略、事实逻辑、内在模式等角度入手，利用特征工程与深度学习技术相结合的方法，从攻防对抗的角度出发，实现社交网络机器生成内容的精准检测. 值得注意的是，本文提及的机器生成文本内容检测与虚假新闻检测[24]既有联系也有区别：(1) 虚假新闻可以从事实逻辑结构进行判断，既可以是人工编造也可以是机器自动生成，可以看作机器生成内容的一种存在形式；(2) 随着深度学习方法的发展，很多社交网络文本内容不再只是局限在虚假新闻生成，而是通过多种手段，生成以假乱真的文本内容，形成“信息轰炸”效应. 本文关注的文本生成内容更侧重该文本是否是由机器生成，其中是否符合事实逻辑只是其中的一个判断标准. 因此，社交网络对抗是指基于特定任务，利用人工智能的方法，自动完成社交网络中内容的生成与检测，实现网络对抗信息的投送与反制，以达到维护网络空间安全的目的.

在线社交网络中内容生成与检测技术的简化流程与关系如图 1 所示. 针对社交网络中特定人群/事件，通过群体/事件感知技术，学习人群/事件内在特征，完成在线社交网络内容的可控自动生成.

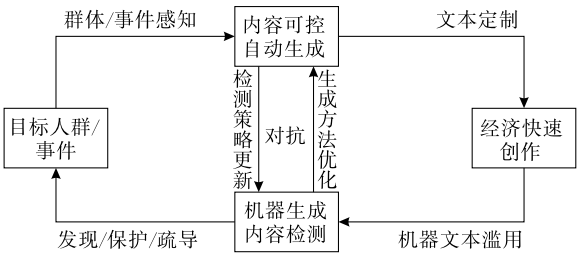


图 1　在线社交网络中内容自动生成和检测技术流程与关系示意图

通过输入相关话题、关键字等信息，自动准确地生成事件布告、新闻速递、产品介绍等定制化文本，可以节省大量的时间与人力资源. 反之，针对由语言模型的易用与经济性等导致的机器生成文本滥用的现象，通过信息传播规律与用户行为分析等方法，对社交网络中的异常信息进行发现，基于对文本质量及内在模式等进行分析与挖掘，对文本内容是否由机器自动成进行检测，进而保护被恶意攻击的目标人群/事件. 其中，内容可控生成方法与机器生成内容检测方法在社交网络对抗过程中，对检测策略与生成模型进行不断地更新与优化，最终达到构建与完善模型的目的.

本文将从相关基础知识、机器生成文本检测模型、文本自动生成模型、已有数据集介绍、未来研究方向及挑战几方面对社交网络对抗技术进行详细阐述. 整体论文组织框架如图 2 所示. 论文中出现的主要符号及说明内容见表 1.

① https://techscience.org/a/2019121801/
② https://openai.com/blog/gpt-3-apps/
③ https://www.forbes.com/sites/forbestechcouncil/2019/01/29/the-80-blind-spot-are-you-ignoring-unstructured-organizational-data/

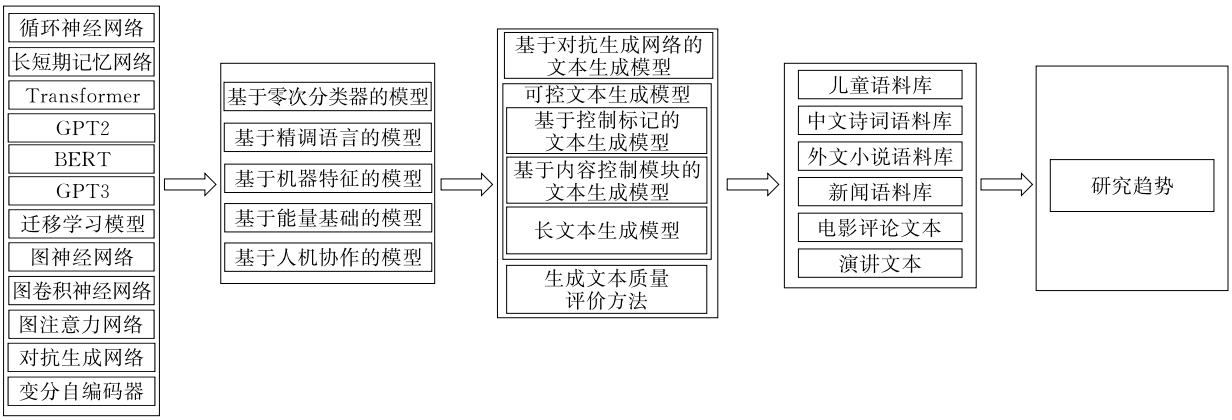


图 2 文章组织框架图

**表 1 符号说明表**

| 英文符号 | 符号说明 |
|---|---|
| $h$ | 隐藏层 |
| $i$ | 输入值 |
| $o$ | 输出值 |
| $\sigma$ | 激活函数 |
| $\boldsymbol{W}$ | 权重矩阵 |
| $b$ | 偏置量 |
| $Q$ | 注意力机制中的查询权重 |
| $K$ | 注意力机制中的键权重 |
| $V$ | 注意力机制中的值权重 |
| $\widetilde{\boldsymbol{A}}$ | 自耦图的邻接矩阵 |
| $\widetilde{\boldsymbol{D}}$ | 自耦图的度矩阵 |
| $\boldsymbol{h}$ | 节点特征 |
| $\alpha_{ij}$ | 注意力系数 |
| $E$ | 能量函数 |
| $KL$ | KL 散度 |
| $JSD$ | JS 散度 |
| $\boldsymbol{s}$ | 句向量 |
| $S_{(j,j-1)}$ | 后句预测分数 |
| $x^+$ | 正样本 |
| $\hat{x}^-$ | 最佳负样本 |
| $sg(\cdot)$ | 梯度停止算子 |
| $a$ | 控制属性 |
| $\mathcal{L}$ | 损失函数 |
| $\boldsymbol{c}$ | 上下文向量 |
| $p$ | 概率分布 |
| $e$ | 码本 |
| $F_{\text{mean}}$ | METEOR 的均值评分 |
| $P$ | 精确率 |
| $R$ | 召回率 |

## 2 相关基础知识

本节简要介绍应用于自然语言处理领域的框架模型，这些模型为自然语言处理的下游任务提供了灵活且有效的模型框架. 例如，本文介绍的在线社交网络对抗技术中的机器生成文本检测技术与社交网络内容自动生成技术等.

**循环神经网络**（**Recurrent Neural Networks**，**RNN**）. 文本上下文之间是有关联的，即上文的输入可以影响下文的输出. 为了使神经网络模型可以应用于文本这样的序列信息，1990 年 Elman[25] 提出了循环神经网络，其神经网络架构如图 3 所示.

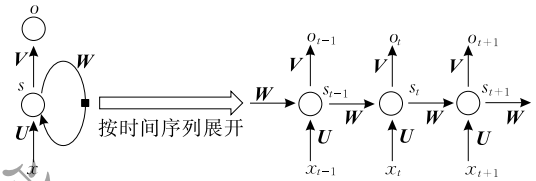


图 3 RNN 架构示意图

RNN 的特点在于其考虑了前一隐藏层 $h_{t-1}$ 的输出 $o_{t-1}$ 对当前隐藏层状态 $h_t$ 的影响，$h_t$ 由 $s_{t-1}$ 层的输出和当前的输入 $x_t$ 共同决定. 其数学表达式如下：

$$o_t=\sigma(\boldsymbol{V}_r h_t) \tag{1}$$

$$h_t=f(\boldsymbol{U}_r x_t+\boldsymbol{W}_r o_{t-1}) \tag{2}$$

其中 $\boldsymbol{U}_r$、$\boldsymbol{V}_r$ 和 $\boldsymbol{W}_r$ 为权重矩阵，分别对应输入到隐藏、隐藏到输出和隐藏到隐藏的连接，$\sigma$ 表示激活函数.

**长短期记忆网络**（**Long Short Term Memory**，**LSTM**）. RNN 的提出解决了神经网络处理序列信息的问题，但当关联信息和当前预测位置之间的间隔较大时，会出现梯度消失的问题，使 RNN 丧失聚合远距离特征的能力. Hochreiter 与 Schmidhuber[26] 于 1997 年提出的 LSTM 可以很好地解决以上远距离信息记忆问题. LSTM 的结构与 RNN 类似，都应用了链式的且输入重复通过相同模块的结构. 不同的是，除隐藏状态 $h_t$ 外，LSTM 添加了有信息存储功能的细胞状态 $c_t$，并且其重复模块不再是单一的神经网络层，而是三个具有不同功能的门结构形成的细胞（cell）. 这三种门结构分别为遗忘门 $f_t$、输入门

$i_t$和输出门 $o_t$. 遗忘门确定从细胞状态中丢失什么信息，

$$f_t=\sigma(\boldsymbol{W}_f[h_{t-1},x_t]+b_f) \tag{3}$$

输入门确定多少信息加入到细胞状态中，

$$i_t=\sigma(\boldsymbol{W}_i[h_{t-1},x_t]+b_i) \tag{4}$$

$$c_t=f_t\odot c_{t-1}+i_t\odot\sigma(W_c[h_{t-1},x_t]+b_c) \tag{5}$$

输出门根据细胞状态确定输出值，

$$o_t=\sigma(\boldsymbol{W}_o[h_{t-1},x_t]+b_o) \tag{6}$$

$$h_t=o_t\odot\tanh(c_t) \tag{7}$$

其中 $\sigma$ 表示激活函数，$\boldsymbol{W}_f$，$\boldsymbol{W}_i$，$\boldsymbol{W}_o$ 分别表示遗忘门、输入门和输出门的权重矩阵，$b$ 为偏置，$\odot$ 表示哈达玛积.

**Transformer 网络.** RNN 模型应用的一大难点是其难以进行并行计算，这大大降低了其运算效率和处理长文本的能力. Transformer[27] 采用了编码-解码架构，即输入一个序列 $(x_1,\cdots,x_n)$，通过编码器将其映射为连续的表示 $z=(z_1,\cdots,z_n)$，再通过解码器依次输出序列 $(y_1,\cdots,y_n)$. 在编码-解码架构中，Transformer 引入了自注意力机制，可以在计算序列语句的表达时综合考虑该序列任意位置的信息，并实现并行计算. 自注意力机制可以由下式表示：

$$attention\_output=\mathrm{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) \tag{8}$$

其中 $\boldsymbol{Q}$ 表示查询向量，$\boldsymbol{K}$ 表示键向量，$\boldsymbol{V}$ 表示值向量，注意力的计算采用了缩放点乘方法，即

$$\mathrm{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{9}$$

其中，$d_k$ 是 $\boldsymbol{K}$ 的维度.

**BERT(Bidirectional Encoder Representation from Transformers).** 已有的语言处理模型可大致分为两类，包括基于文本特征的模型和基于微调参数的模型. 基于特征的模型通过预训练获得提取文本特征的能力，在进行下游任务时直接使用训练好的模型提取文本特征，将所提取特征传递给下游处理；微调参数的模型通过预训练初始化模型参数，在应对不同的任务时使用迁移学习的方式针对下游任务的实际需求对模型参数进行调整. 两类方法的共性特点在于它们在预训练过程中都采用了单向的语言模型，因此不能够对上下文信息进行充分利用. Devlin 等人[3] 提出 BERT 模型采用了新的预训练目标函数，使预训练过程中可以使用双向的语言模型进行训练，从而提高模型对上下文信息的利用能力. 同时，BERT 模型还对语言处理模型可处理的文本级别进行了扩充，从原有的单字级别和词级别扩充到了句子级别，使模型能够理解句子之间的关系，从而处理更长的文本.

**GPT-2.** 为了避免因为在单领域数据集上进行单任务训练，导致语言模型对其它种类的任务缺乏适用性，Radford 等人[4] 提出了基于无监督学习的语言预训练模型 GPT-2. GPT-2 基于 Transformer[27] 模型结构，并且在很大程度上和 GPT 模型相似，但同时也做出了改进. 它将层归一化移至每一子块的输入部分，在最后的自注意力模块后添加了归一化层. 另外，残差层的参数初始化也做出了改变，可以根据网络深度进行调节. GPT-2 在包含多个不同领域语言样本和不同任务方向的 WebText 数据集[27] 上进行了训练，在多个语言处理任务方面都取得了优异的表现.

**GPT-3.** 2020 年 Brown 等人[5] 提出了一种新的语言模型 GPT-3. 该模型继续采用了单向语言模型的训练方式，但将模型参数扩大到了 1750 亿，并采用了 45TB 数据进行训练. 这一模型的核心是解决自然语言处理模型的通用性问题，其主要目标是在采用更少的邻域数据情况下，使模型即使不经过参数精调也可以解决更多其他种类的问题. 在许多自然语言处理数据集上，针对翻译、问答、文本填空以及需要即时推理和领域适应的任务，GPT-3 均表现出了出色的性能.

**迁移学习(Transfer Learning).** 在传统的机器学习模型中，经训练集训练后得到的模型往往只能解决特定的问题. 遇到不同的问题时，由于数据分布的差异，或是标注数据过期，会导致训练好的模型无法被扩展使用. 为了增强模型在存在相关性的大部分数据和任务中的广泛应用能力，迁移学习的概念被提出. Pan 和 Yang[28] 给出了迁移学习概念的数学定义，即根据原领域 $D_S=\{X_S,f_S(X)\}$ 和原学习任务 $T_S$，利用迁移学习得到目标领域 $D_T=\{X_T,f_T(X)\}$ 的预测函数 $f_T(\cdot)$，其中 $D_S\neq D_T$ 或 $T_S\neq T_T$.

**图卷积神经网络(Graph Convolutional Networks, GCN).** 传统卷积神经网络中的基本算子依赖于数据的平移不变性[29]，即只适用于欧几里得结构数据，无法在图数据这样的非欧数据上应用. 2013 年 Bruna 等人[30] 首次提出基于谱方法的图上的卷积神经网络，但因其较高的复杂度难以应用. 2016 年 Kipf 等人[31] 提出 GCN，利用图卷积的一阶近似形式参数化卷积核，大大降低了模型的复杂度. GCN 的层间传播规则如式(10)，

$$\boldsymbol{H}^{(l+1)}=\sigma(\widetilde{\boldsymbol{D}}^{-\frac{1}{2}}\widetilde{\boldsymbol{A}}\widetilde{\boldsymbol{D}}^{-\frac{1}{2}}\boldsymbol{H}^{(l)}\boldsymbol{W}^{(l)}) \tag{10}$$

其中拉普拉斯算子由包含自耦的无向图的邻接矩阵

$\widetilde{\boldsymbol{A}}$ 和度矩阵 $\widetilde{\boldsymbol{D}}$ 提前计算得出，初始输入图的顶点特征矩阵，训练时只需更新每层网络的权重矩阵 $\boldsymbol{W}^{(l)}$. 这种方法使 CNN 可应用于图数据，并在引用网络和知识图谱等数据集上的分类精度和计算效率有明显的提升.

**图注意力网络(Graph Attention Networks，GAT)**. GCN 中的卷积算子依赖于图的结构，只能针对特定结构训练并应用于相似结构的图数据. 2017 年 Veličković 等人[32]提出了 GAT，可对单个节点特征 $\boldsymbol{h}_i$ 及其相邻节点特征 $\boldsymbol{h}_j$ 进行独立分析，不需要整张图的信息. GAT 的注意系数可由式(11)计算，

$$\alpha_{ij}=\frac{\exp(\mathrm{LeakyReLU}(\boldsymbol{a}^{\mathrm{T}}[\boldsymbol{W}\boldsymbol{h}_i\|\boldsymbol{W}\boldsymbol{h}_j]))}{\sum_{k\in\mathcal{N}_i}\exp(\mathrm{LeakyReLU}(\boldsymbol{a}^{\mathrm{T}}[\boldsymbol{W}\boldsymbol{h}_i\|\boldsymbol{W}\boldsymbol{h}_k]))} \tag{11}$$

其中 $\boldsymbol{a}^{\mathrm{T}}$ 表示单层前馈网络的权重向量，$\boldsymbol{W}$ 表示对特征向量进行线性变换的网络权重矩阵，$\mathcal{N}_i$ 为节点 $v_i$ 的邻居节点集.

文章中使用了 softmax 和 LeakyReLU 函数对注意系数进行归一化和激活处理. 得到注意系数后，用其计算相应节点的输出特征向量，

$$\boldsymbol{h}_i'=\sigma\Big(\sum_{j\in\mathcal{N}_i}\alpha_{ij}\boldsymbol{W}\boldsymbol{h}_j\Big) \tag{12}$$

将注意力机制用于图神经网络，可对各个节点进行并行运算，适用于多种结构的图，达到增强了模型的灵活性与鲁棒性的目的.

**变分自编码器(Variational Auto-Encoder，VAE)**. 在生成问题中，难以度量生成数据和真实数据的分布及二者之间的差异. 为解决这一问题，2013 年 Kingma 等人[33]提出了变分自编码器，将问题转化为用变分的方法近似目标分布的均值和标准差优化. 与传统的自编码器不同，作者令编码器 $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ 的输出为隐变量 $\boldsymbol{z}$ 分布的均值 $\boldsymbol{\mu}$ 和标准差 $\boldsymbol{\sigma}$，采样 $\boldsymbol{z}$ 后输入到解码器 $p_\theta(\boldsymbol{x}|\boldsymbol{z})$ 生成数据. 其中编码器网络参数为 $\boldsymbol{\phi}$，解码器网络参数为 $\boldsymbol{\theta}$，则单个数据点的边际似然为

$$\log p_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)})=KL(q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}^{(i)})\|p_\theta(\boldsymbol{z}|\boldsymbol{x}^{(i)}))+\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\phi};\boldsymbol{x}^{(i)}) \tag{13}$$

训练时只需最大化似然函数的变分下界 $\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\phi};\boldsymbol{x}^{(i)})$，

$$\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\phi};\boldsymbol{x}^{(i)})=-KL(q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}^{(i)})\|p_{\boldsymbol{\theta}}(\boldsymbol{z}))+\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}^{(i)})}[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}|\boldsymbol{z})] \tag{14}$$

$$\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\phi};\boldsymbol{x}^{(i)})\simeq\frac{1}{2}\sum_{j=1}^{J}(1+\log((\boldsymbol{\sigma}_j^{(i)})^2)-(\boldsymbol{\mu}_j^{(i)})^2-(\boldsymbol{\sigma}_j^{(i)})^2)+\frac{1}{L}\sum_{l=1}^{L}\log p_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}|\boldsymbol{z}^{(i,l)}) \tag{15}$$

式(14)中第一项为关于编码器输出分布和目标分布(文献中假设为标准正态分布)的 $KL$ 散度(Kullback-Leibler divergence)，第二项为给定编码器时生成数据的期望，两项结合同时保证了生成数据的准确性和多样性. 经作者证明，此变分下界如式(15)表示，其中 $J$ 表示向量 $\boldsymbol{\mu}$ 和 $\boldsymbol{\sigma}$ 的元素数量，$L$ 表示一个数据点的采样数量，满足更新参数必须的可微性.

**生成对抗网络(Generative Adversarial Network，GAN)**. 同样为了解决数据分布表示的问题，2014 年 Goodfellow 等人[34]提出了基于生成器与判别器对抗博弈学习训练的神经网络 GAN，其网络结构如图 4 所示. 模型中生成器 $G$ 捕获真实数据分布，判别器 $D$ 判断输入来自训练样本的概率，根据 $D$ 输出结果的交叉熵的变式指导模型的学习.

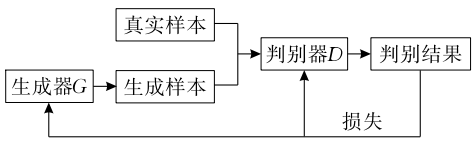


图 4 生成对抗网络结构

$$\begin{aligned}\max_D V(D,G)&=\max_D(\mathbb{E}_{\boldsymbol{x}\sim p_{\mathrm{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})]+\mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1-D(G(\boldsymbol{z})))])\\&=-\log(4)+2JSD(p_{\mathrm{data}}\|p_g)\end{aligned} \tag{16}$$

$$\min_G V(D,G)=E_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1-D(G(\boldsymbol{z})))] \tag{17}$$

训练时，分别依据式(16)和式(17)优化 $D$ 和 $G$，其中 $p_{\mathrm{data}}$ 为训练样本的分布，$p_z$ 为噪声变量(生成器输入)的先验分布，$p_g$ 为生成样本的分布，$\boldsymbol{x}$ 为真实数据，$\boldsymbol{z}$ 是生成器 $G$ 的输入，$JSD$ 是 $JS$ 散度(Jensen-SHannon divergence). GAN 不需要复杂的马尔科夫链，免去了损失函数设计的困难，相比于其他生成模型，GAN 可以更加完美地学习到训练样本的分布.

## 3 社交网络机器生成文本检测方法

本节从文本分布、文本特征、精调模型、能量模型等几方面对机器生成文本检测方法进行了详细总结，分析了各类方法生成在线社交网络文本内容的适用场景与存在的不足. 在线社交网络文本检测方法对比总结内容见表 2.

### 3.1 基于零次分类模型的检测方法

目前主流的文本生成任务中，多采用 GPT 或 GROVER 作为预训练模型. GPT 系列模型基于 WebText 数据集进行训练，该数据集包含了 Reddit① 上得到 3 个点赞以上的文章，其规模达到了 40 GB. GROVER 采用 RealNews 数据集作为训练的语料

① https://www.reddit.com/

**表 2　在线社交网络文本检测方法对比总结**

| 文本检测模型 | 应用场景 | 模型优点 | 模型缺点 |
|---|---|---|---|
| 基于零次分类模型 | 针对推特、微博等在线社交网络的短文本生成内容的统计学检测方法 | 简单易行，节省计算资源 | 需要预知语言模型模型和参数设置，且统计学特征的有效性需进一步检验 |
| 基于精调预训练语言模型 | 针对社交网络特定文本类型的检测方法 | 在特定类型数据集上的表现较好 | 模型表现依赖于训练数据中负样本的质量，泛化能力较差 |
| 基于机器特征的检测模型 | 广泛对抗各类型社交网络机器文本 | 挖掘机器文本的统计特征与逻辑缺陷，模型泛化能力强，有助于进一步改进语言模型 | 对机器特征如何产生的解释仍不完善，无法从语义角度对文本进行检测 |
| 基于能量基础的检测模型 | 检测不同类型的社交网络机器文本 | 跨模型结构检测的鲁棒性强 | 过于依赖负样本的多样性，由单一语料库训练的模型泛化能力较弱 |
| 基于人机协作的检测机制 | 检测短社交文本，提高人类识别机器文本的能力 | 结合人类和机器的优势，捕捉文本中的语义错误和分布规律 | 在大规模应用过程中人力成本较高 |

库，其内容是网络新闻文章. 这些基于社交网络文本的预训练模型，可以生成强互动性、口语化、甚至包含事实知识的高质量社交网络文本，危害网络空间安全. 此外，社交网络文本通常具有不规范性，即用户在写作时较为随意，文本内容间的逻辑关系较弱，因此在微博文本、推特文本等短文本社交网络内容场景中，较难捕捉文本的结构特征和逻辑特征. 但文本的统计特征不受文本规范性的限制，可以做为分辨机器文本和人类文本的重要特征. 因此，针对以上预训练模型与社交网络文本内容特点，基于零次分类模型的检测方法能够完成在线社交网络文本生成内容检测，该小节对此类方法进行了总结与介绍.

在零次分类器中，使用了预训练语言模型（如 GPT-2、GROVER）来检测其自身或是相似模型生成的文本. 零次分类器不需要使用标注的数据对模型进行进一步的训练，其检测流程如图 5 所示. 检测器采用白盒设置，即假设生成器所用的模型及各项设置已知. 通过待检测文本和生成器所用模型产生的分布特征的相似度来对文本进行检测.

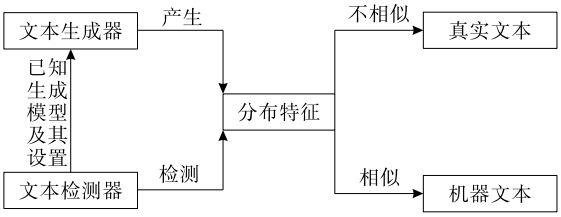


图 5　基于零次分类模型的文本检测流程

2019 年 OpenAI 的工作[42]提出了零次分类模型的简单基准，该工作使用 GPT-2 模型作为检测器，对 GPT-2 自身生成的文本进行检测. 其策略是用 GPT-2 计算文本的总对数概率，并设置概率阈值，对数概率高于该阈值的文本被认为是假文本，低于该阈值的文本则认为是真文本. 该工作使用该基准模型对由两种不同的采样方式（随机采样和 Top-$K$ 采样，其中 $K=40$）生成的文本进行检测和分析. 在检测使用随机采样方法生成的文本时，随着检测器参数量的增大，检测的准确度相应提高. 参数量为 1557M 的 GPT-2 作为检测器，检测参数量为 124 M 的 GPT-2 生成的文本，其准确度达到了 95.7%. 但参数量为 124M 的 GPT-2 检测器，检测参数量为 1557M 的 GPT-2 生成的文本时，准确度仅为 76.0%. 在检测使用 Top-$K$ 方法生成的文本时，检测模型参数量对模型的检测效果并没有显著提升，检测模型参数量的提升有时甚至会造成检测准确率的下降. 但大多数的实验表明，检测准确率多在 81.1%上下浮动. 以上述方法为基准，许多工作借助概率统计的思想设计零次分类模型，大幅提高了对机器文本的检测准确率. 2019 年 Gehrmann 等人[35]提出一种基于采样策略的文本检测方法，命名为 GLTR（Giant Language model Test Room）①. Hashimoto 等人[36]认为可以基于文本与模型的相似性来区分真实文本和机器生成文本. 受其启发，GLTR 提出文本的分布模式是判断文本是否由机器自动生成重要依据的基本思路. 在基于神经语网络的语言模型中，模型通常根据语句序列中位置靠前的词语来对下一个词语的可能性分布进行编码. 解码策略则依据一定的规则在这样的分布中决定选择哪一个词语. 很多生成语言模型都采用了基于概率分布的解码策略，即基于概率从输出的分布中对候选词进行采样. 但是，当输出的概率分布中大量存在低概率的词语时，这样的词语整体也会占据相当大的被采样的可能性，从而导致生成的文本出现语意矛盾等问题. 这些错误很容易使得人们识别出该文本是由机器自动生成.

为了避免这样的情况出现，许多研究改进了从模型输出的分布 $x_t \sim q(x_t | x_1, \cdots, x_{t-1}, p_\theta)$ 中采样的

① https://github.com/HendrikStrobelt/detecting-fake-text

方法，即随机解码方法. 传统的采样方法对采样不做任何限制，被称为纯采样(pure sampling). 采用这种方法容易采样到分布尾部的词元，即低可能性词元，所以会使生成的文章不通顺，容易被检测出来. 为了解决纯采样可能产生的问题，需要限制采样范围，即从整体候选词中选择出部分 $W \subset V$，再进行采样. Top-$K$ 采样(核采样)[6]和 Top-$P$ 采样[37]是目前常用的两种采样方法. Top-$K$ 采样将采样范围限制在前 $k$ 个可能性最高的词元中，其采样范围 $W$ 满足 $\max\left(\sum_{x\in W} p_\theta(x|x_1,\cdots,x_{t-1})\right)$. 这种方法的缺点是，其采样范围 $K$ 是一个固定值，这导致同一个值针对不同文本时的采样效果不能保证达到最优. Top-$P$ 采样方法克服了这一缺陷，其思想是定义一个阈值 $P\in[0,1]$，在一个所有词元的可能性之和不小于 $P$ 的最小集合 $W$ 中进行采样，即 $\sum_{x\in W} p_\theta(x|x_1,\cdots,x_{t-1})\geq P$. 这样采样范围就可以随着文本的不同进行变化，生成更加自然流畅的文本.

虽然以上采样方式可以有效骗过人工审查，但是其固定模式很容易被基于特征分析的机器学习方法识别出来. GLTR 是一种可视化工具，它采用白盒设置，即使用 GPT-2 模型对由 GPT-2 自身生成的文本进行检测. 由于检测器与生成器采取了相同的模型，检测器可以被视为预知了生成文本的概率分布. 通过计算文本词汇在模型中的概率密度，可以确定该词汇是否采样于一个范围有限的单词集. 机器生成的文本受限于采样策略，文本中的词语采样于模型预测出现可能性较高的单词集，而人类作品则不受概率因素限制. 其具体测试方法包括测试下一个词语与预测词语相近的可能性，

$$p_{\text{det}}(X_i=\hat{X}_i|X_{1:i-1}) \tag{18}$$

测试下一个词元的可能性级别，

$$p_{\text{det}}(X_i|X_{i:1}) \tag{19}$$

和测试分布的熵值，

$$-\sum_{w} p_{\text{det}}(X_i=w|X_{1:i-1}) \tag{20}$$

前两个测试方法用来检测生成的词是否采样于概率分布中的高概率词语，最后一个测试方法用来检测测试文本是否满足预设分布. 结果发现机器生成文本更倾向于选择那些高可能性词汇，这也意味着机器生成文本的选词空间受到限制. 在 GLTR 的辅助下，未经过训练的用户分辨机器生成文章的准确率从 54%上升至 72%. 该方法的有效性被 Ippolito 等人[38]的实验证明. 他们发现，当机器生成的文本行文流畅，足以欺骗人类时，其采样策略必须是从模型计算出的高概率词语中按照一定的策略进行选择. 这样的文本就会被 GLTR 检测出来. 当模型不以概率为依据随机采样词语生成文本时，其生成的文本不够通顺合理，容易被人类辨识.

然而，GLTR 采用了白盒设置，这需要预知假文本是由何种模型生成的，实际应用场景有限. 后续的一些研究发现，通过零次分类进行跨语言模型的检测也是可行的.

2020 年 Adelani 等人[20]发现基于 GPT-2 的精调模型可以生成流畅且令人难以辨别的虚假评论. 他们基于和 GLTR 相似的零次分类策略，使用 GROVER 作为检测器计算词汇出现的概率，以此作为判断评论真假的依据. 实验结果表明 GROVER 作为检测器在亚马逊评论数据集上可以达到 97% 的检测准确率.

但也有研究质疑零次分类模型的有效性. 2020 年 Schuster 等人[39]通过随机增删文章段落中的否定词，发现基于统计学的检测方法并不能有效地分辨出文章内容的可靠性. 根据实验结果，他们认为判别器仅仅利用文本的分布特征来进行判别是不够的，必须建立起事实核查模型.

由此可见，零次分类模型尽管简单易行，但其过于依赖文本的统计特征，白盒设置这一前提在实际在线社交网络对抗场景中也较难满足. 因此，需要进一步完善零次分类模型，提高其泛化能力及事实核查能力.

### 3.2 基于精调预训练语言模型的检测方法

社交网络文本的内容具有多样性，包括新闻类长文本、微博类短文本、评论类交互文本等. 尽管预训练模型采用的语料库十分庞大，如 GPT-2 采用 40 GB 的文本数据训练，BERT 的训练数据包含 33 亿个单词，Grover 的语料库达到 120 GB，但是对于多样化的社交网络文本，仍无法在各种类型的文本数据集上都取得出色的检测效果. 只有针对特定的文本检测任务，基于特定小规模数据集对预训练模型进行参数精调，方能达到文本内容精准检测的目的，其基本流程如图 6 所示.

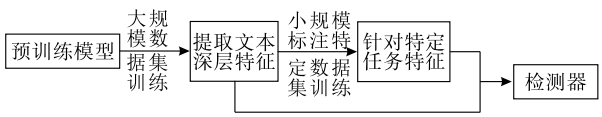


图 6 基于精调预训练语言模型的机器生成文本检测流程

2019 年 Zellers 等人[18]基于文本生成与检测对抗的思想，提出了 Grover 模型. 该方法是通过训练

大规模 Transformer 语言模型获得，并且设计了三种不同规模大小的模型. 最小的模型(Grover-Base)是基于预训练语言模型 GPT 和 BERT-Base 的，具有 12 层 Transformer 网络和 1.24 亿个参数；中级模型(Grover-Large)是基于预训练语言模型 BERT-Large 的，具有 24 层 Transformer 网络和 3.55 亿个参数；最大的模型(Grover-Mega)是基于预训练语言模型 GPT2 的，具有 48 层 Transformer 网络和 15 亿个参数.

在具体的机器生成文本检测的过程中，作者通过在 Grover 顶层加入了一个线性分类器来实现对文本进行检测的功能. 具体做法是在文章末尾加入[CLS]标记并提取出标记处的隐藏状态，该状态作为线性层的输入即可对文章进行分类，即检测文章的真实性. 在判别器大小相同的情况下，针对机器生成的文本检测问题，Grover 检测的准确率高于 BERT、GPT-2 和 FastText[40] 等方法. 对于使用 Grover 生成的文本，人类能够识别的准确率只有 73%，而使用 Grover 自身检测准确率可达到 92%. 作者还对该模型检测文本的机制进行了探究，认为曝光偏置(Exposure Bias)和方差降低方法(Variance Reduction)产生了能够被检测出的缺陷. 然而 2020 年 Uchendu 等人[41] 通过不同的文本生成方法设置对已有方法有效性进行验证，发现 Grover 存在在检测其他语言模型生成的文本时表现不佳的情况.

针对 Zellers 等人[18] 提出的“检测生成文本的最有效工具是生成模型本身”这一论断存在诸多争论，部分学者认为这一言论限制了检测模型的泛化能力. 2019 年 Solaiman 等人[42] 采用基于精调 RoBERTa 的文本检测方法检测由 1554M GPT-2 生成的文本. RoBERTa 模型为非生成式模型，且采用了掩码，它和 GPT-2 模型架构完全不同. 但在检测 GPT-2 生成文本的任务中，基于精调 RoBERTa 的检测模型达到了 95%的准确率，其表现优于精调后的 GPT-2 模型. 此外，该研究发现精调时使用的负样本内容会影响检测器的检测准确率. 当使用混合了不同采样方法生成的负样本数据集和随机长度文本的负样本数据集对 RoBERTa 进行训练后，检测器的泛化能力得到了提升，在不同采样方法和文本长度的测试集中均取得了较高的准确率.

2020 年 Fagni 等人[43] 实验发现 RoBERTa 在检测机器生成的推特时效果远优于传统的机器学习模型和神经网络模型. 随后的研究也发现了 RoBERTa 在检测不同模型生成的文章[41] 和 GPT-2 生成的产品评论[20] 时效果也是最出众的. 这些研究表明 RoBERTa 在公开数据集上精调训练后的具有极强的可扩展能力.

精调预训练语言模型针对在线社交网络中多样化文本类型的检测提出了有效的解决方案，但精调语言模型的检测效果依赖于精调训练数据的选择. 若精调所用的训练数据种类单一(如负样本基于同一种采样方法生成)，或质量较差(如负样本使用参数量较小的模型生成)，会极大影响检测器的表现. 因此需要进一步挖掘生成文本中隐藏的深层机器特征，以减少检测模型对于训练数据的依赖，增强检测器的泛化能力.

### 3.3　基于机器特征的检测方法

随着文本生成技术的发展，生成模型已不再仅仅局限于大型的预训练模型，越来越多的新型文本生成模型(如 CTRL[44]、FVN[45])被应用于创作社交网络文本. 因此，对于检测方法的研究不应局限于对抗大型预训练模型，而需要挖掘机器生成文本的方式和逻辑与人类的差异，以达到广泛对抗社交网络机器生成文本的目的. 许多工作对文本生成模型产生的特征进行了研究，并以此为依据进行文本识别，其基本检测流程如图 7 所示.

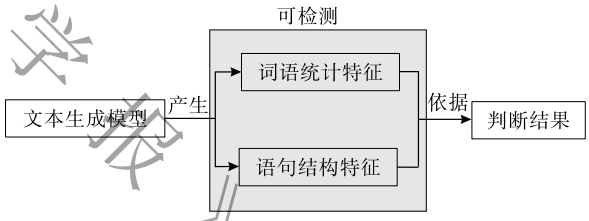


图 7　基于机器特征的机器生成文本检测流程

2019 年 OpenAI 的工作[42] 测试了最简单的检测模型——基于一元语法和二元语法词频-逆文档频率(TF-IDF)特征的逻辑回归检测器对虚假文本的检测效果. 当该检测器检测 GPT-2 使用随机采样的方法采样生成的文本时，检测准确率在 74%(1554M GPT-2)到 88%(124M GPT-2)范围内. 当该检测器检测 GPT-2 使用 Top-$K$($K=40$)的方法采样生成的文本时，检测准确率在 93%(1554M GPT-2)到 97%(124M GPT-2)范围内. 该实验结果说明 Top-$K$ 采样方法生成的文本中易出现可以被基于一元语法和二元语法词频-逆文档频率捕捉的特征.

2020 年 Tay 等人[46] 的工作对不同的机器生成文本策略进行了逆向工程，细致研究了机器生成文本中出现的缺陷，为上述工作提供了解释. 该工作对

使用不同采样策略和采样参数(如 Top-$K$ 和 Top-$P$ 采样方法中的 $K$ 和 $P$)生成的文本进行分类.实验结果表明,即便使用最为简单的词袋模型与线性分类器,其分类准确率也可以几乎达到随机判断的两倍.这一实验结果表明采样策略会使机器生成的文章出现大量可检测的特征,且不同的采样策略导致文章出现可检测特征的数量也大不相同.检测器对基于 Top-$P$ 采样生成的文章进行分类时达到的准确率高于对 Top-$K$ 采样生成的文章分类达到的准确率.这一结果说明采用 Top-$P$ 采样策略生成的文本蕴含的可检测特征多于 Top-$K$ 采样生成的文本.有趣的是,该工作发现基于词袋模型的检测器和基于序列信息编码(如 LSTM)的检测器对该文本分类任务达到的准确度相似.这一结果说明机器生成文本产生的可检测特征是有关写作风格的特征(如词语的选择),而不是时序上的远距离依赖特征(如词序).这一实验结果和 Uchendu 等人[41]的实验结论一致.

2020 年 Zhong 等人[47]从篇章一致性的角度出发,综合了词语选择、词频与词序特征,构建了实体一致数(Entiy Consistency Count,ECC)和语句一致数(Sentence Consistency Count,SCC)这两个指标.其中,ECC 指在下一个语句窗口中重复提到的实体数,SCC 是指在下一个语句窗口中重复提到相同实体的语句数.作者分别计算语料库中真假文本的 ECC 与 SCC,发现机器生成文本中 ECC 与 SCC 数值均小于真实文本.据此,作者提出了利用文本的事实结构进行机器生成文本检测的方法.在具体实施过程中,作者使用 RoBERTa[48]进行词语表达,利用 AllenNLP 实验室开源的最新的命名实体识别工具箱①进行实体分析,并最终使用图模型抽取文章的事实结构.其中,图模型的节点代表实体,节点之间的边代表实体关系,利用知识图谱进行初始化.对实体的特征表达采用了基于维基百科的实体嵌入表达方式(Wikipedia2Vec)[49],并采用多层图卷积神经网络[[31]生成了结合知识库的句段表达.多层神经网络可以由下式表达为

$$\boldsymbol{H}_e^{(i+1)}=\sigma(\widetilde{\boldsymbol{D}}^{-\frac{1}{2}}\widetilde{\boldsymbol{A}}\widetilde{\boldsymbol{D}}^{-\frac{1}{2}}\boldsymbol{H}_e^{(i)}\boldsymbol{W}_i) \tag{21}$$

其中 $\boldsymbol{H}_e^{(i)}$ 是由第 $i$ 层神经网络生成的节点 $e$ 的表达,$\widetilde{\boldsymbol{D}}^{-\frac{1}{2}}\widetilde{\boldsymbol{A}}\widetilde{\boldsymbol{D}}^{-\frac{1}{2}}$ 是图 $G$ 的拉普拉斯算子,$\boldsymbol{W}_i$ 是第 $i$ 层的权重.基于多层神经网络节点生成的句段表达可由下式计算:

$$\boldsymbol{y}_i=\frac{1}{N_i}\sum_{j=0}^{N_i}\sigma(\boldsymbol{W}_S\boldsymbol{H}_{i,j}+\boldsymbol{b}_s) \tag{22}$$

其中 $\boldsymbol{y}_i$ 是句子的表达,$\boldsymbol{W}_S$ 是权重矩阵,$\sigma$ 是激活函数,$\boldsymbol{H}_{i,j}$ 是句子 $i$ 中第 $j$ 个节点的表示,$\boldsymbol{b}_s$ 是偏置量.

最后作者采用序列模型和后句预测模型(Next Sentence Prediction,NSP)对句子之间的关系进行建模,并计算出整个文本的表达以用于最后的判断.

$$\boldsymbol{D}=\sum_{j=1}^{s}S_{(j-1,j)}\times[\tilde{\boldsymbol{y}}_{j-1},\tilde{\boldsymbol{y}}_j] \tag{23}$$

其中 $\boldsymbol{D}$ 是最终整篇文本的表达,$S_{(j-1,j)}$ 是两个句子相连的可能性分数.通过在新闻风格和网页文本风格的数据集上进行训练和测试,发现该利用文本事实结构的模型检测精确度优于 GPT-2、BERT 和 XLNet[50]等基于 Transformer 的模型.

2020 年 Jawahar[51]提出基于篇章一致性的模型[52]对机器生成文本进行检测.他们认为现存的检测方法只将文本看作是词语的序列,而忽视了文中语句结构的关系.这种方法的核心思想是假设由人类创造的文本句间是有联系的,而机器生成文本则不然.作者分别使用了实体网格模型(Entity Grid,EGRID)[53]、实体图模型(Entity Graph,EGRAPH)[54]、语句平均模型(Sentence Averaging,SENTANG)[52]和段落序列模型(Paragraph Sequence,PARSEQ)[52]对分别采用 Pure 采样、top-$k$ 采样以及 top-$p$ 采样的机器生成文本进行检测,并将实验结果与使用 NGRAM、BERT 和 GPT-2 的分类模型分别进行对比.结果显示尽管采用 BERT 和 GPT-2 这样的预训练模型检测的精确度较高,但篇章一致性模型不需要提前得知生成器的内部表示,而且其对结果的可解释性较强,并可以达到令人满意的精确度.

对生成模型产生的机器特征的研究极大提高了检测模型未来的泛化能力,但在线社交网络文本(如网络新闻)是具有事实内容的文本.文本内容的真实性也是检测机器文本的重要特征.基于机器特征的模型无法在语义层面上理解文本,不能检测出文本中存在的事实错误.

### 3.4 基于能量基础模型的检测方法

如前所述,在线社交网络文本内容形式多样,涉及领域广泛,这对应用于社交网络文本的检测模型的泛化性提出了高要求.近期的一些研究采用能量基础模型进行文本检测,此类方法对模型的泛化能力提升较大.能量基础模型的示意图如图 8 所示.

① https://demo.allennlp.org/named-entity-recognition

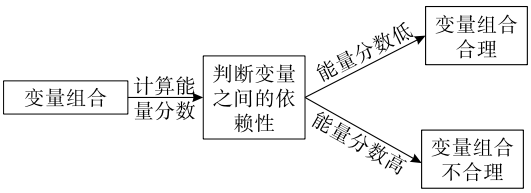


图 8　能量基础模型原理示意图

基于统计学的机器学习模型的主要目的是对变量之间的依赖性进行编码. 通过捕获这些依赖关系，可以基于已知变量值使用模型来回答相关未知变量值的问题. 能量模型[55]通过将能量值与变量的每个取值进行关联来捕获他们之间的依赖性. 在进行预测或决策的过程中，通过设置观察变量的值，最终找到使能量最小化的其余变量的值.

在最通用的模型中，观测值 $X$ 做为输入量，对应和 $X$ 最匹配的模型输出值为 $Y$. 能量模型的最终目的就是找出所有可能的 $Y^* \in \mathcal{Y}$，使得 $E(Y,X)$ 取值最小，即

$$Y^* = \arg\min_{Y \in \mathcal{Y}} E(Y,X) \tag{24}$$

该模型可以解决预测、分类、决策、排序、检测等问题.

基于能量模型的思想，2019 年 Bakhtin 等人[56]借助现在强大的预训练模型 Transformer[27]、Gated CNN[57]和 GPT-2[4]直接生成负样本来训练基于能量基础的判别器. 作者将目标能量函数设置为 $E(w_1,\cdots,w_n|c;\theta)$，达到计算给定上下文 $c$ 的输入词元(tokens)序列 $w_1,\cdots,w_n$ 和参数设定 $\theta$ 的联合兼容度(joint compatibility)的目的. 学习的目的就是获得能量低的词元序列，比如人工创作的文本会获得更低的能量. 以此为目标，作者采用交叉熵损失来训练能量函数，

$$\mathcal{L}_{\mathrm{BCE}} = -\log(\sigma(-E(x^+|c;\theta))) + \log(\sigma(-E(\hat{x}^-|c;\theta))) \tag{25}$$

其中 $x^+$ 是正样本，$\hat{x}^-$ 是最佳负样本，$\sigma$ 是 sigmoid 函数. 作者对该模型分别进行了跨语料库(训练集和测试集来自不同的语料库)和跨模型结构(训练集和测试机的负样本由不同语言模型生成)测试，实验结果表明能量基础模型在跨模型结构测试中表现出了较强的鲁棒性. 但是在跨语料库测试中，由单一语料库训练的模型泛化能力较弱，不能很精准地检测出由其他语料库训练出的虚假文本. 当检测器的训练数据囊括待检测的所有语料库时，该能量基础模型在各语料库上的测试精准度会有显著提升.

### 3.5　基于人机协作的检测方法

多数社交网络文本的篇幅较短，如微博文本篇幅被限制在 140 字以内，商品评论往往只有几句话以表达顾客的态度. 短文本写作任务对生成模型来说相对简单，较短的文本长度可以避免大量可检测机器特征(如高可能性词语出现频率)的产生，为文本检测工作带来困难. 人类在评判时往往会注意到文章的矛盾之处或语义错误[38]，这有效地弥补了机器检测对深层语义理解的不足的短板，可以极大提高短文本的检测准确率. 另一方面，机器则更善于计算文本的分布规律，这一优势可以很好地针对长文本，如网络新闻，进行检测，而人类往往无法辨别这些潜在的统计规律. 因此，在检测工作中建立人机协作的机制，使两者在检测过程中发挥各自的长处，可以扩大社交文本检测的范围，提高整体的检测准确率. 图 9 展示了人机协作机制的检测流程.

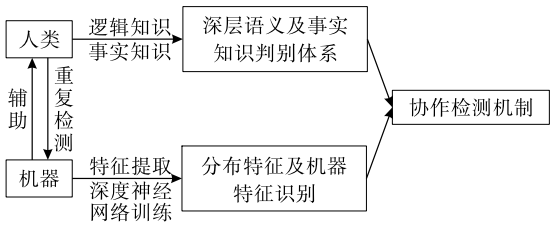


图 9　基于人机协作的机器生成文本检测方法示意图

基于以上思想，2020 年 Dugan 等人[58]提出了一种名为 RoFT(Real or Fake Text tool)的文本检测工具. 这个工具以游戏的形式邀请用户完成标记工作. 这个游戏包含几轮，但是每一轮开始呈现给用户的句子都确保是人写的而不是机器生成的. 接下来用户可以选择展示更多的句子，并对这些句子进行判断，确定这些句子中大多数是否还是由人写的. 当用户确定展示的句子中大多数句子都是机器生成的，这个测试就结束. 同时，这里的句子叫做可以区分人类创作与机器生成内容的分界线句子.

该工具可以逐句展示文本，其展示的文本会从真实文本变化为机器生成文本，而用户在判断的过程中需要对自己的决定给出相应的说明. 如果人们判断的真实文本与生成文本的分界线距真实的分界线较远，则说明机器成功地骗过了人类. 这一工具的一大创新点是增加了注释功能，人们可以记录下自己作出判断的原因. 当答案揭晓后，人们可以通过自己的注释来反思自己的决策并找到生成文本和真实文本潜在的不同. 这可以有效地训练人们区分真实文本和生成文本的能力. 这种人机结合的思路也为机器生成文本检测提供了新思路.

人机协作机制的概念较为新颖，但人类判断的主观性，不可预测性较强，缺乏公认的判别标准，该

方法的有效性还有待验证.另外,社交网络中存在海量文本内容,若人机协作机制大规模应用于社交网络文本在线检测任务,会极大增加检测工作的人力成本,有违社交网络文本自动检测技术的初衷.

本节对社交网络机器生成文本检测方法进行了较为全面的总结,详细介绍了各类优秀的检测方法在社交网络对抗中的应用场景.但是,由于在线社交网络文本内容的特殊性,已有方法仍存在较大缺陷,亟须改进和提升,以应对纷繁复杂的网络空间安全问题.

## 4 社交网络文本自动生成方法

本节将从基于生成对抗网络的文本生成方法、可控文本生成方法以及生成文本质量评价三方面对当前相关工作进行总结.此外,针对在线社交网络对抗特点,对每一类方法的使用场景与不足进行了总结.

### 4.1 基于生成对抗网络(GAN)的文本生成方法

不可控的文本生成无法直接应用在社交网络中模拟用户发表文字内容,但它是可控生成、长文本生成的基础,是二者在句层面上语义通顺、语法正确的前提.因此,提高不可控文本生成模型的性能也是不容忽视的.

在众多不可控文本生成模型中,基于 GAN 的生成模型性能突出.这得益于其对抗的结构,在无监督的生成任务中表现良好.但其仅仅被设计用于连续的实值数据[59],无法使用离散数据(如自然语言等)对网络参数进行细微的调整.针对以上特点,一些研究将对抗的思想与常用于文本处理的方法结合,在文本生成方面取得了进步.相关工作的基本思想和类别总结如图 10 所示.

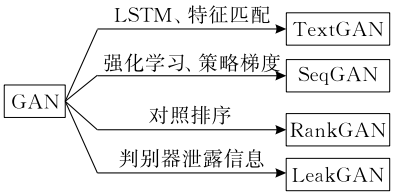


图 10 基于 GAN 的文本自动生成方法类别

2016 年 Zhang 等人[60]尝试将对抗的结构用于文本生成,并提出了名为 TextGAN 的文本生成模型.该模型以 LSTM 作为生成器,CNN 作为判别器,采用特征匹配的方法代替原 GAN 中的优化目标.实验表明,TextGAN 的生成器可以很好地模拟真实数据的分布(数据特征的期望和协方差),生成的句子基本上达到了语法正确、语义合理,表现出一定的流畅性的要求,但存在隐空间中部分区域的措辞选择不准确和句子结构不稳定的缺点.

2016 年 Yu 等人[61]将 GAN 扩展应用到离散词元序列上,提出了序列生成对抗网络(Sequence GAN,SeqGAN).在处理离散数据时,传统 GAN 的生成器难以传递梯度更新参数,判别器无法评估不完整序列以指导生成过程.针对以上难题,作者使用强化学习与 GAN 相结合以解决上述两个问题.具体包括利用蒙特卡洛搜索方法补全每一次动作对应的各种完整序列,输入到判别器产生奖励,通过策略梯度的方法将奖励传回生成器更新参数,其流程如图 11 所示.作者将 SeqGAN 分别在中文四行诗集①、奥巴马演讲②和诺丁汉音乐集③上进行训练并生成文本序列,与已有工作相比,SeqGAN 具有显著的优势.尤其是在诗歌创作方面,SeqGAN 的性能与人类创造的真实数据相当.

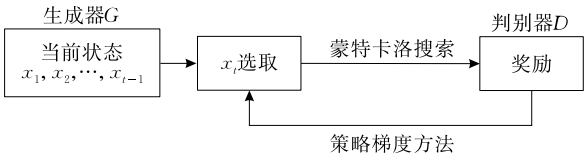


图 11 SeqGAN 基本计算流程

2017 年 Lin 等人[62]针对传统的 GAN 模型难以处理离散数据和只能进行二分类判别两个缺点,提出了 RankGAN 模型.该模型的核心思想是将策略梯度方法用于解决不可微的问题.将原有的判别器改为排序器,并引入反例作为对照组,根据模型生成文本和真实文本之间的排序信息训练该模型,放宽了二分类的限制.作者分别将 MLE(Maximum Likelihood Estimation)、SeqGAN 和 RankGAN 在中文诗集[63]、COCO[64]和莎士比亚剧集[65]上进行训练,使用 BLEU 和人为评估对生成的文本进行打分.实验结果显示,在以上三个代表性的数据集上,RankGAN 性能明显优于 MLE 和 SeqGAN.

GAN 判别器输出的判别结果通常隐含信息较少,且无法在生成过程中对文本结构进行指导,导致其难以生成较长文本.针对此问题,2017 年 Guo 等人[66]提出了 LeakGAN,令判别器向生成器泄漏高维特征信息提供指导.模型中判别器分为两层,首先根据当前的句子 $s_t$ 抽取特征 $f_t=\mathcal{P}(s_t)$,抽取出的特征经降维映射后通过 sigmoid 函数输出判别结果,同时向生成器泄露进行指导.为了令生成器充分利用判别器泄露的高维信息,作者参考了 Vezhnevets

① http://homepages.inf.ed.ac.uk/mlap/Data/EMNLP14/
② https://github.com/samim23/obama-rnn
③ http://www.iro.umontreal.ca/~lisa/deep/data

等人[67]提出的 FeUdal 网络的分层结构，将生成器分为基于 LSTM 模型的 MANAGER 模块和 WORKER 模块，分别接收 $f_t$ 和 $s_t$，综合依据指导信息和当前状态生成新单词. 在不同长度的生成文本和真实文本上的实验表明，与之前的解决方案相比，LeakGAN 获得了更高的 BLEU 和人类评分，这种优势在生成长句子时尤为显著.

在线社交网络网络对抗中的内容生成通常是针对特定群体、特定主题实施的. 根据不同的需求，不同的内容发布平台，社交网络文本内容也需要定制. 尽管基于 GAN 的文本生成模型在提升文本流畅度方面效果显著，但用户无法精准控制生成文本的内容，这导致基于 GAN 的文本生成模型难以直接应用于社交网络对抗场景.

### 4.2　可控文本生成

在线社交网络对抗中的文本往往都有具体的含义，比如对事件的叙述、对物体的描述、特定观点与情感的表达等. 这些文本不仅需要保证通顺流畅，还要包含特定的内容、立场、情绪等属性，对生成模型的要求更高. 可控文本生成模型可以对生成的文本的属性进行控制，如话题控制、情感控制、时态控制等，令生成的文本在社交网络中拥有更广的应用范围. 表 3 对目前效果较好的可控文本生成模型及重要特征进行了简要总结.

**表 3　可控文本生成模型特征总结**

| 文本生成模型 | 前置内容 | 应用场景 | 训练集 | 算法性能 |
|---|---|---|---|---|
| CTRL[44] | 控制代码及文本 | 符合控制代码话题的文本 | 控制代码（如网络链接）及真实文本 | 文本很好地符合控制代码的主题 |
| Grover[18] | 新闻元数据（如话题、标题、日期、作者）或新闻主体 | 缺失的元数据内容或新闻主体 | 从 RealNews 中获得的新闻文章及其元数据 | 由元数据调整，经大型训练集训练后的 Grover 模型的困惑度仅为 8.7. 人工评测中的可信度得分高于真实文本 |
| PPLM[68] | 文章开头 | 后续文章 | 无再训练或精调 | 人类评估下，生成文本话题相关性高于 CTRL 与 WD(Weighted Decoding)[69]，情感准确性高于 CTRL、GPT2-FT-RL 和 WD |
| CoCon[70] | 文章开头及其真实后续 | 重建的后续文章 | 无再训练或精调 | 重构相似度方面与 GPT-2 比较，BLEU-4、NIST-4[71] 和 METEOR[72] 评估分数分别提高 0.22、7.09 和 6.14. 话题相关度和情感控制方面，CoCon 高于 GPT-2、PPLM 和 CTRL |
| FVN[45] | 内容标记和风格控制标记 | 符合确定内容/风格的文本 | 带有内容和风格标注的真实文本 | 在 METEOR 和 ROUGE-Ld[73] 的评估下优于 CVAE[74] 等基准算法，符合预期的内容和风格 |

#### 4.2.1　基于控制标记的方法

控制文本生成内容的方法之一是利用作者、创作日期和文章来源等元数据生成控制标记，将控制标记拼接在输入的文本序列数据前，使得生成模型在训练时能够学习到元数据与文章内容间的关联.

2017 年 Gupta 等人[75]认为传统的 VAE 与 RNN 相结合虽然可以自由生成文本，但不适合对给定句子进行多种改写. 针对以上问题，作者提出了一种基于 VAE 与 LSTM 相结合的可控文本生成模型. 训练数据以句子对（原始句 $\boldsymbol{s}^{(o)}$ 和改写句 $\boldsymbol{s}^{(p)}$ 的形式给出，输入端的两个 LSTM 编码器将 $\boldsymbol{s}^{(o)}$ 转变为 $\boldsymbol{x}^{(o)}$，再接收 $\boldsymbol{x}^{(o)}$ 将 $\boldsymbol{s}^{(p)}$ 转变为向量 $\boldsymbol{x}^{(p)}$，其输出通过前馈神经网络产生 VAE 编码器的平均值和方差参数. 输出端将 $\boldsymbol{x}^{(o)}$ 作为 LSTM 解码器的初始输入，将隐变量 $\boldsymbol{z}$ 输入到解码器的每一阶段，共同影响改写句的生成. 该种方法的变分下界定义为

$$\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\phi};\boldsymbol{x}^{(p)},\boldsymbol{x}^{(o)})=\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}^{(o)},\boldsymbol{x}^{(p)})}[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}^{(p)}|\boldsymbol{z},\boldsymbol{x}^{(o)})]-KL(q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}^{(o)},\boldsymbol{x}^{(p)})\|(\boldsymbol{z})) \quad (26)$$

在保证 $\boldsymbol{z}$ 的后验和先验接近的前提下，通过使式(26)的函数最大化，可以得到期望的 $\boldsymbol{x}^{(p)}$. 作者将模型在 MSCOCO[64] 和 Quora 数据集上进行测试，并使用 BLEU、METEOR 和 TER(Translation Error Rate)[76] 评测. 实验结果表明，在没有任何超参数调整的情况下，该模型性能显著优于当时已有的基于 VAE 模型的生成算法.

2019 年 Keskar 等人[44]发现，尽管 GPT-2 和 BERT 等预训练模型已经有了生成高质量文本的能力，但是如果不能与生成文本内容的控制规则相结合，已有技术就很难实现特定任务内容自动生成. 针对以上问题，他们提出了可以进行可控文本内容生成的模型 CTRL. 其核心思想是在语言模型中加入控制信息 $c$，通过对语料库的数据进行分类，在每个具体序列的内容前都加入了其类型描述，从而使语料与类型之间产生联系. 其采用的语言模型可表述为

$$p(x|c)=\prod_{i=1}^{n}p(x_i|x_{<i},c) \quad (27)$$

其中，$x=\{x_1,\cdots,x_n\}$ 表示给出的样本序列，$x_i\in x$

是一个固定长度的符号集合.经过训练学习,包含 $n$ 个词元(tokens)的简单样本序列 $x$ 被嵌入到组成一个序列的 $n$ 个对应向量中.每一个嵌入的向量包含了学习到的词元(tokens)的嵌入和位置嵌入的总和.序列经过表示后被输入到 $l$ 层注意力网络中,其中每一层由两个模块构成.第一个模块是 $k$ 头的注意力模块,定义为

$$\text{Attention}(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z})=\text{softmax}\left(\frac{\text{mask}(\boldsymbol{X}\boldsymbol{Y}^T)}{\sqrt{d}}\right)\boldsymbol{Z} \tag{28}$$

$$\text{Multihead}(\boldsymbol{X},k)=[h_1;\cdots;h_k]\boldsymbol{W}_o \tag{29}$$

其中,$\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z}$ 分别表示键向量,查询向量和值向量,$h_j=\text{Attention}(\boldsymbol{X}\boldsymbol{W}_j^1,\boldsymbol{X}\boldsymbol{W}_j^2,\boldsymbol{X}\boldsymbol{W}_j^3)$.

第二个模块是采用激活函数 ReLU[77] 的前馈神经网络.

$$FF(\boldsymbol{X})=\max(0,\boldsymbol{X}\boldsymbol{U})\boldsymbol{V} \tag{30}$$

其中,$\boldsymbol{U}$ 和 $\boldsymbol{V}$ 是对应定义维度的参数.

每个模块都先将输入进行层归一化[78-79]计算,然后采用残差网络[80]进行连接.最终,在最后一层输出每个单词的分数,实现文本的可控生成.值得注意的是,该模型在解码器采样过程中,使用了惩罚采样方法进行采样,通过减少前面生成词语的分数,进行接近贪婪算法的采样,从而避免了采样过程中出现的词语重复问题[37].

2019 年 Zellers 等人发布了 Grover[18] 模型,该模型主要针对新闻内容进行控制生成.由于一篇完整的新闻包括发布网站、日期、作者、题目和新闻正文五个部分,因此在 Grover 中,选择使用这五部分信息的联合分布作为最终生成的新闻内容的概率分布,如下所示:

$$p(\text{domain},\text{date},\text{authors},\text{headline},\text{body}) \tag{31}$$

在生成过程中,Grover 要求给定五个部分中的几个部分作为上下文,然后根据上下文先生成目标部分,再将生成的部分添加入上下文中,并对原有给定的部分进行调整.举例来说,当给定了发布网站、日期和题目后,尝试使用 Grover 完善新闻的其它内容,那么 Grover 会先使用给定的三个部分作为上下文,生成正文部分的内容.在生成正文后,正文部分也会被作为上下文内容的一部分,Grover 此时利用四个部分作为上下文,再对作者部分进行生成.完成作者部分的生成后,Grover 会使用发布网站、日期、作者和正文作为上下文,对题目部分进行重新生成,从而找到和正文内容更加贴合的文章标题.生成标题后,一篇受发布网站、日期和题目控制的 Grover 生成新闻就完成了.

通过上面举例可以看出,该模型采用的是自左至右的语言模型,即根据前 $x_{i-1}$ 个词语来预测第 $x_i$ 个词语.当选定五方面的内容后,如何从这五方面的联合概率分布中准确采样是一个难点.若是采用固定的采样顺序,则会因边际效应过高而阻止某些控制条件的采样;如果不规定控制条件的采样顺序,模型就需要学习处理 $|\boldsymbol{F}|!$ 种潜在的采样顺序,其中 $\boldsymbol{F}$ 表示控制条件.针对以上问题,Grover 采用了一种新的采样方法,为每方面约束都规定了特定的开始和结束符号.如果要开始生成某一方面的内容,先在内容中加入这方面的开始符号,然后从模型中采样,直到采样到该方面的结束符号.这样就不需要定义每方面内容生成的先后顺序,也不用预先学习处理各种可能会出现的采样顺序.实验结果发现上述五方面信息的完整性越高,训练集越大,生成内容的困惑度越低.同时当拥有相同数据量时,Grover 的困惑度比 GPT-2 低 5.作者猜测可能是 GPT-2 的训练集中有非新闻的文章,造成对模型生成文本的质量存在影响.在人类辨识的实验中,人们辨认出 Grover 模型生成的新闻的准确度只有 73%.这在证明了 Grover 模型的强大功能的同时,也为人们可能被机器生成的文本所欺骗敲响了警钟.

2020 年 Shu 等人[45]提出了一种名为聚焦变分网络(Focused-Variation Network,FVN)的可控文本生成模型.该方法通过学习每个属性的码本(codebook)间不相交的隐空间来达到可控生成的多样性和流畅性.训练过程中,输入目标内容 $c$(内含 $g$ 个标记槽)、目标风格 $s$ 和参考文本 $t$,其中 $t$ 分别通过内容转换和风格转换编码器生成隐向量 $\boldsymbol{z}^C$ 和 $\boldsymbol{z}^S$.在码本 $\boldsymbol{e}^C$ 和 $\boldsymbol{e}^S$ 中将所有隐向量量子化[81]后记录为 $[\boldsymbol{e}_1^C,\cdots,\boldsymbol{e}_K^C]$ 和 $[\boldsymbol{e}_1^S,\cdots,\boldsymbol{e}_N^S]$,并选出与当前隐向量最接近的 $\boldsymbol{e}_k^C$ 和 $\boldsymbol{e}_n^S$;$c$ 和 $s$ 分别通过内容和风格编码器生成隐向量 $\boldsymbol{v}^C$ 和 $\boldsymbol{v}^S$,将 $\boldsymbol{v}^C,\boldsymbol{v}^S,\boldsymbol{e}_k^C,\boldsymbol{e}_n^S$ 四个向量输入到基于 LSTM 的解码器中生成文本.解码器的损失是每个单词的交叉熵损失之和,即

$$\mathcal{L}_{\text{Dec}}-\sum_l \log P(t_l') \tag{32}$$

为保证生成的文本传达正确的内容和风格,将生成文本的向量表示 $\boldsymbol{o}_L$ 输入到内容和风格解码器中进行反向预测,得到的结果与输入的 $c$、$s$ 做对比.其相应的损失函数为

$$\mathcal{L}_{\text{CTRL}}=-\sum_g \log F^C(\boldsymbol{e}_k^C)-\log F^S(\boldsymbol{e}_n^S)-\sum_g \log F^C(\boldsymbol{z}'^C)-\log F^S(\boldsymbol{z}'^S) \tag{33}$$

其中 $F$ 为分类函数,$\boldsymbol{z}'^C$ 和 $\boldsymbol{z}'^S$ 分别为 $\boldsymbol{o}_L$ 关于内容和风格的向量表达.

为了令生成文本与参考文本的向量表示存在于同一空间内，$\boldsymbol{o}_L$需要在词嵌入的量子化码本$\boldsymbol{e}^V$中选择. 则三个码本对应的 VQ 损失[81]为

$$\mathcal{L}_{\mathrm{VQ}}^{C}=\left\|sg(\boldsymbol{z}^{C})-\boldsymbol{e}_{k}^{C}\right\|_{2}^{2}+\beta^{C}\left\|\boldsymbol{z}^{C}-sg(\boldsymbol{e}_{k}^{C})\right\|_{2}^{2} \tag{34}$$

$$\mathcal{L}_{\mathrm{VQ}}^{S}=\left\|sg(\boldsymbol{z}^{S})-\boldsymbol{e}_{n}^{S}\right\|_{2}^{2}+\beta^{S}\left\|\boldsymbol{z}^{S}-sg(\boldsymbol{e}_{n}^{S})\right\|_{2}^{2} \tag{35}$$

$$\mathcal{L}_{\mathrm{VQ}}^{V}=\left\|sg(\boldsymbol{o}_{L})-\boldsymbol{e}_{v}^{V}\right\|_{2}^{2}+\beta^{V}\left\|\boldsymbol{o}_{L}-sg(\boldsymbol{e}_{v}^{V})\right\|_{2}^{2} \tag{36}$$

其中 $sg(\cdot)$是停止梯度算子.

综上，该模型训练过程中需要最小化的损失为

$$\mathcal{L}=\mathcal{L}_{\mathrm{Dec}}+\mathcal{L}_{\mathrm{CTRL}}+\mathcal{L}_{\mathrm{VQ}}^{C}+\mathcal{L}_{\mathrm{VQ}}^{S}+\mathcal{L}_{\mathrm{VQ}}^{V} \tag{37}$$

生成过程中，依据训练过程的方法得到$\boldsymbol{v}^C$和$\boldsymbol{v}^S$，再从与内容和风格相关的训练数据中获得$\boldsymbol{z}^C$和$\boldsymbol{z}^S$并映射到码本中索引最接近的向量得到$\boldsymbol{e}_k^C$和$\boldsymbol{e}_n^S$，以使用训练出的解码器生成文本. 作者将 FVN 在 PersonageNLG[82]和 E2E[83]数据集上进行测试，结果表明，FVN 表现出了最好的性能. 根据评估人员的评价，该模型生成的文本可与正确标注的文本相媲美.

基于控制标记的可控文本生成模型可以在一定程度上控制文本生成内容. 但由于控制标记与训练数据间的关系是隐式的，生成模型仅可以学习到写作风格、文章内容等较高层面的文本信息，无法对文章进行句子级、甚至是字符级的细粒控制.

### 4.2.2　基于属性分类器的文本生成方法

基于属性分类器的文本生成方法是近年新兴的内容控制方法. 它将一个或多个属性分类器嵌入 GPT-2 等预训练模型，通过属性分类器控制生成文章的性质，使文本生成模型可以达到对内容的严格控制.

2018 年 Peng 等人提出了一个分析现有故事语料库的框架[84]，可以根据给定的故事结尾类型(快乐或悲伤)和故事情节可控地生成故事. 该框架主要包含了监督分类器和关键字提取器进行数据分析，并最终用条件 RNNs 进行文本生成.

在结尾类型控制的方法中，作者用一个基于双向 LSTM 的逻辑回归分类器[85]来分析和提取现有故事的信息，提取的信息池化后用 softmax 分类得到感情标签. 生成器是一个条件语言模型，将情感标签用嵌入矩阵映射为一个向量，结合故事情节共同生成结尾. 在故事情节控制的方法中，作者采用 RAKE 算法[86]进行关键字提取. 在每个句子中提取出最重要的单词作为故事情节，其生成器也是一个条件语言模型，用双向 LSTM 将故事情节词编码成一个向量，然后依据条件概率解码生成故事.

对受控和不受控模型根据 ROCstories 数据集[87]计算困惑度，受控模型的困惑度低于不受控模型. 其中受结尾类型控制的模型由于只有一个比特的信息，与不受控模型的差异较小. 在人工评估下，受控模型生成的故事与给定信息的匹配度更高，流畅度也优于不受控模型.

2018 年 Hu 等人[88]提出了一种结合了 VAE 和整体属性识别器的模型，可以实现对生成模型的语义结构进行控制. 这个模型解决了文本数据不连续和隐变量中各属性纠缠的问题. 核心思想是采用了一个新的变量$\boldsymbol{c}$与 VAE 中的隐变量$\boldsymbol{z}$结合，其中$\boldsymbol{c}$关联着句子中想要控制的特定属性，而$\boldsymbol{z}$则控制着其他属性. 通过将联合分布$(\boldsymbol{z},\boldsymbol{c})$生成的句子作为判别器的输入实现. 对于 VAE 和判别器的优化则采用一种 wake-sleep 交替优化算法[89].

在训练的过程中，当生成器每一步取特定单词作为输出时，单词间的离散性会使判别器的优化过程无法进行. 为了解决这个问题，作者将生成器 $G$ 每一步的输出改为 softmax 概率值，使其变成一个连续可导的量，即

$$\begin{aligned}\hat{\boldsymbol{x}}\sim G(\boldsymbol{z},\boldsymbol{c})&=p_G(\hat{\boldsymbol{x}}\mid\boldsymbol{z},\boldsymbol{c})\\&=\prod_t p(\hat{\boldsymbol{x}}_t\mid\hat{\boldsymbol{x}}^{<t},\boldsymbol{z},\boldsymbol{c})\end{aligned} \tag{38}$$

其中，$\hat{\boldsymbol{x}}=\{\hat{x}_1,\cdots,\hat{x}_T\}$是词元序列，$\hat{\boldsymbol{x}}^{<t}$表示 $t$ 时刻前的词元. 这些离散的词元是从一个 softmax 分布函数中采样获得，如

$$\hat{\boldsymbol{x}}_t\sim\mathrm{softmax}(\boldsymbol{o}_t/\tau) \tag{39}$$

其中，$\boldsymbol{o}_t$表示 softmax 函数输入的对数向量，$\tau>0$ 用来将温度归一化为 1. 然后将温度逐渐进行“退火”，从而使连续分布逼近离散分布. 作者为训练编码器和生成器定义了三种损失函数，分别是标准 VAE 的损失函数，式(40)，判别器 $D$ 的损失函数，式(41)，即生成的句子与希望的$\boldsymbol{c}$的分布之间的交叉熵函数，和为保证属性独立性而定义的隐变量损失函数，式(42). 令$\boldsymbol{\theta}_G$，$\boldsymbol{\theta}_E$分别代表生成器 $G$ 和编码器 $E$ 的参数，$\boldsymbol{x}$是给出的观测值，则三个损失函数定义如下：

$$\begin{aligned}\mathcal{L}_{\mathrm{VAE}}(\boldsymbol{\theta}_G;\boldsymbol{\theta}_E;\boldsymbol{x})=&KL(q_E(\boldsymbol{z}\mid\boldsymbol{x})\,\|\,p(\boldsymbol{z}))-\\&\mathbb{E}_{q_E(\boldsymbol{z}\mid\boldsymbol{x})q_D(\boldsymbol{c}\mid\boldsymbol{x})}\left[\log p_G(\boldsymbol{x}\mid\boldsymbol{z},\boldsymbol{c})\right]\end{aligned} \tag{40}$$

$$\mathcal{L}_{\mathrm{Attr},\boldsymbol{c}}(\boldsymbol{\theta}_G)=-\mathbb{E}_{p(\boldsymbol{z})p(\boldsymbol{c})}\left[\log q_D(\boldsymbol{c}\mid\widetilde{G}_\tau(\boldsymbol{z},\boldsymbol{c}))\right] \tag{41}$$

$$\mathcal{L}_{\mathrm{Attr},\boldsymbol{z}}(\boldsymbol{\theta}_G)=-\mathbb{E}_{p(\boldsymbol{z})p(\boldsymbol{c})}\left[\log q_E(\boldsymbol{z}\mid\widetilde{G}_\tau(\boldsymbol{z},\boldsymbol{c}))\right] \tag{42}$$

训练编码器和生成器的损失函数即为三者的加

权和，即

$$\min_{\theta_G}\mathcal{L}_G=\mathcal{L}_{\text{VAE}}+\lambda_c\mathcal{L}_{\text{Attr},c}+\lambda_z\mathcal{L}_{\text{Attr},z} \tag{43}$$

而判别器则是使用半监督学习方法进行训练，即采用少量有属性标签的样本和大量生成器生成的无标签样本进行训练，以达到精准预测句子属性和评价根据特定隐式编码的恢复的特征的错误. 最后，作者在实验中基于情感和时态两个属性做了测试，并以 S-VAE 作为比较的对象，结果发现该模型判别器对情感的分类效果在准确度上有了 3%～4% 的提升. 除此之外，当控制语句时态时，模型生成的句子也可以保持时态不同而句意基本不变. 但是该模型只能生成长度不超过 15 词的句子，在文本丰富性上还有很大的提高空间.

2019 年 Dathathri 等人发现可控文本生成模型通常是通过对预训练模型进行精细调节，或是从零开始训练一个大型语言模型得到的. 这些大规模模型对数据及配置环境存在依赖性，移植要求较高. 针对以上难题，作者受到计算机视觉领域中 PPGN[90] 方法的启发，提出了一个“即插即用”的模型 PPLM[68]. PPLM 模型分为语言模型 $p(x)$ 和条件属性模型 $p(a|x)$ 两部分，其中语言模型用来生成文本，属性模型用来使生成文本具有特定的属性.

PPLM 通过属性模型计算出来的梯度更新语言模型中的隐藏层，从而使每一次循环中隐藏层得到的概率分布都更为接近目标属性，其更新定义为

$$\Delta H_t\leftarrow\Delta H_t+\alpha\frac{\nabla_{\Delta H_t}\log p(a|H_t+\Delta H_t)}{\|\nabla_{\Delta H_t}\log p(a|H_t+\Delta H_t)\|^{\gamma}} \tag{44}$$

其中，$H_t$ 是由截止时间 $t$ 所有 Transformer 的 key 和 value 对构成，$\alpha$ 是步长，$\gamma$ 是 Transfomer 中归一化项的比例系数.

同时为了保证文本的流畅性，PPLM 采用 KL 散度和规范几何平均的方法来指导语言模型训练学习. 其中采用规范几何平均方法的目的是将生成分布和语言模型绑定. 文章表示了这样的“即插即用”模型可以用来进行“文本排毒”，即消除文章中的有冒犯性的语言使用.

2020 年 Chan 等人提出的名为 CoCon[70] 文本生成模型，可以实现对文本内容进行词汇和短语级的精确控制. 该模型的核心思想是通过在 GPT-2 的编码器和解码器中间加入一个 CoCon 模块，将目标内容加入编码文本的表示中. CoCon 模块定义为

$$\boldsymbol{h}'_{:,t-1}=\text{CoCon}(\boldsymbol{h}^{(c)}_{:,l_c},\boldsymbol{h}_{:,t-1}) \tag{45}$$

其中 $\boldsymbol{h}^{(c)}_{:,l_c}=\text{enc}(c)$ 是目标内容的表示，$l_c$ 是文本序列的长度. CoCon 模块是基于 Transformer 构建的. 在将文本的编码输入 CoCon 模块前，需要将目标内容的 $\boldsymbol{K}^{(c)},\boldsymbol{V}^{(c)}$ 拼接到初始的注意力矩阵 $\boldsymbol{K},\boldsymbol{V}$ 中，

$$\boldsymbol{K}'=[\boldsymbol{K}^{(c)};\boldsymbol{K}] \tag{46}$$

$$\boldsymbol{V}'=[\boldsymbol{V}^{(c)};\boldsymbol{V}] \tag{47}$$

$$\boldsymbol{A}=\text{softmax}(\boldsymbol{Q}\boldsymbol{K}'^{\text{T}})\boldsymbol{V}'=\text{softmax}(\boldsymbol{W})\boldsymbol{V}' \tag{48}$$

其中 $\boldsymbol{A}$ 是注意力矩阵，$\boldsymbol{W}$ 是注意力的权值. CoCon 模块的输出由一个位置前馈层计算为

$$\boldsymbol{h}'_i=FF(\boldsymbol{a}_i) \tag{49}$$

$$\tilde{\boldsymbol{o}}_t=\text{dec}([\boldsymbol{h}_{:,t-2};\boldsymbol{h}'_{t-1}]) \tag{50}$$

$$p_{\theta,\psi}(x_t|\boldsymbol{c},x_{:,t-1})=\text{softmax}(\tilde{\boldsymbol{o}}_t) \tag{51}$$

其中 dec 表示解码器，$\theta,\psi$ 分别代表 CoCon 模块出和语言模型的参数. 实现经过内容调整的文本表示送入解码器进行文本生成.

作者采用了一种自监督学习的方式来训练 CoCon 模块，通过将一段给定的文本序列 $x=\{x_1,\cdots,x_{t-1},x_t,\cdots,x_l\}$ 分成两段 $[x^a;x^b]$，其中 $x^a=\{x_1,\cdots,x_{t-1}\},x^b=\{x_t,\cdots,x_l\}$. CoCon 可以通过将 $x^b$ 作为条件输入，即 $c=x^b$，当作目标控制内容，可以获得期望输出 $x^a$. 为了达到以上训练目的，作者定义了自重建损失，式(52)、空文本损失，式(53)、循环重建损失，式(54)以及对抗损失，式(55)这四种损失函数.

$$\mathcal{L}_{\text{self}}=-\sum_{i=t}^{l}\log p_{\theta,\psi}(x_i|(c=x^b,\{x_1,\cdots,x_{i-1}\})) \tag{52}$$

$$\mathcal{L}_{\text{null}}=-\sum_{i=t}^{l}\log p_{\theta,\psi}(x_i|(c=\varnothing),\{x_1,\cdots,x_{i-1}\}) \tag{53}$$

$$\mathcal{L}_{\text{cycle}}=-\sum_{i=t}^{l}\log p_{\theta,\psi}(y_{\text{cycle}}=x^b|(c=y_{x,x'}),(p=x^a)) \tag{54}$$

$$\mathcal{L}_{\text{adv}}=\mathbb{E}_x[\log f_{\text{disc}}(enc(x))]+\mathbb{E}_y[\log(1-f_{\text{disc}}(enc(y))] \tag{55}$$

其中 $f_{\text{disc}}$ 表示判别器网络，该网络用来判断输出的文本表示是否由 CoCon 生成，$enc$ 表示编码器，最终达到了仅以 $x^a,x^b$ 为输入数据完成 CoCon 模块自监督训练的效果. 实验结果证明了加入 CoCon 模块的语言模型对生成文本的内容可以做到精确到词和短语的控制，而且也可以在话题和文本情感这样宏观的方面对文章内容进行控制. CoCon 所具有的模块化性质也让它可以应用于其它可控文本生成器中，如 PPLM、CTRL 等.

可控文本生成的方法逐渐多元化、精准化，但其整体结构和生成机制决定了其在生成较长的多句子

文本上的劣势. 顶层逻辑指导的缺乏和粗粒度层面联系的忽略导致了可控文本生成逻辑不连贯、前后文信息缺失等问题，令其难以用于在线社交网络对抗中新闻、故事等长文本内容生成场景.

### 4.3　长文本生成

在线社交网络中的文本内容包含大量的长文本，如新闻、广告等，往往由多个句子组成. 除可控性和流畅性外，还要求文本在整体上逻辑顺畅、主题一致. 针对此特点，长文本生成模型逐渐能够学习到更高层次的语义信息，提高了在社交网络中的应用程度.

2018 年 Yao 等人[91]提出了一种分层生成框架 Plan-and-Write，通过规划故事线的方法约束故事的逻辑和合理性. 作者使用一系列关键词表示故事线，令每个关键词与故事中每个句子一一对应，并且给出了两种生成策略：动态规划和静态规划.

动态规划中，关键词 $l_i$ 和故事句子 $s_i$ 交叉生成，如图 12 虚线，二者的条件概率表示为

$$p(l_i|[\boldsymbol{t},\boldsymbol{s}_{1:i-1}],l_{i-1}),p(s_i|[\boldsymbol{t},\boldsymbol{s}_{1:i-1}],l_i) \tag{56}$$

其中 $\boldsymbol{t}$ 表示故事题目，$\boldsymbol{s}_{1:i-1}$ 表示全部上文. 静态规划中，先根据题目利用 Seq2Seq 模型生成完整的故事线，再用故事线指导故事生成. 静态规划相比前者写作灵活性降低，但可以增强故事的连贯性与逻辑性. 作者将没有规划机制的相似方法 Inc-S2S 和 Cond-LM 作为基线算法分别与动/静态方法进行比较，在重复率评估和人工评估下，两种新方法的性能均优于基线算法，其中静态方法的性能最佳.

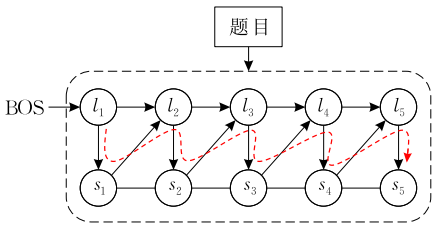


图 12　动态规划流程

文本生成是少信息到多信息的映射，若仅有先验和训练数据，会使生成的文本质量降低，这种现象在长文本生成时尤为显著. 为了解决这一问题，2020 年 Guan 等人提出了一种将外部知识整合进预训练模型的生成方法. 作者将外部常识知识库 ConceptNet 和 ATOMIC 中的三元组类型信息按照固定模板转换为可读的句子，然后使用这些句子按照自左向右方法对已经经过预训练的 GPT-2 进行后训练. 为了进一步捕捉句子之间的逻辑关系，作者尝试使用多任务对模型进行辅助训练. 任务要求模型将合理的故事从逻辑混乱、主题无关或情节重复的故事中区分出来，结合在数据集上的语言建模训练，其损失函数为

$$\mathcal{L}_{\mathrm{ST}}=\mathcal{L}_{\mathrm{LM}}+\lambda\mathcal{L}_{\mathrm{CLS}} \tag{57}$$

其中 $\mathcal{L}_{\mathrm{LM}}$ 为语言建模的损失，$\mathcal{L}_{\mathrm{CLS}}$ 为分类任务的损失. 作者使用 ROCStories 数据集对模型进行评估，实验表明，外部知识和多任务学习有效地提升了模型的性能，令其优于 ConS2S 和 Fusion 等基线方法.

虽然隐式地编码有利于外部知识和网络参数的整合，但仍缺乏对如何使用知识的具体指导. 对此，2020 年 Xu 等人[92]提出了一种显式地将规划机制和外部知识运用到生成模型的方法 MEGATRON-CNTRL，其框架由关键词预测器、知识检索器、知识排序器和条件文本生成器组成.

作者将整体框架设计为动态规划. 首先根据上文为下一句预测一组关键词，然后根据关键词在外部知识库中检索相关的三元组，并使用模板转换为知识句子，最后将这些句子按照同语境的相关性进行排序，并将上文和排名最高的知识句子作为生成下一个句子的条件. 重复上述过程至生成结束.

由于人类用于控制生成的知识句子无法接触，用于知识排序器训练的真实标记数据集难以获得. 作者利用 RAKE 算法提取文本中的关键词，再利用 USE[93]算法找出在知识库里检索出的句子中与语境相关性最高的部分，作为训练排序器的伪真实数据. 在 ROCStories 数据集上的实验结果表明，MEGATRON-CNTRL 的性能优于 Plan-and-Write[91]和知识强化 GPT-2[92]，证明了规划机制和外部知识对长文本生成的积极作用.

2019 年 Koncel-Kedziorski 等人[94]提出了一种基于知识图谱进行长文本生成的模型 GraphWriter. 作者在摘要文本上使用 SciIE 构建知识图谱数据集 AGENDA，并将其转换为无标记的联通二部图作为指导生成的先验知识. 对于知识图谱和文章题目，作者分别使用图 Transformer 和双向 RNN 进行编码，然后通过注意力层计算出关于图谱和题目的上下文向量 $\boldsymbol{c}_g$ 和 $\boldsymbol{c}_s$. 解码成文本的过程中，作者引入了复制机制，即首先根据隐向量 $\boldsymbol{h}_t$ 和上下文向量 $\boldsymbol{c}_t=[\boldsymbol{c}_g\|\boldsymbol{c}_s]$ 计算出复制输入信息的概率

$$p=\sigma(\boldsymbol{W}_{\mathrm{copy}}[\boldsymbol{h}_t\|\boldsymbol{c}_t]+b_{\mathrm{copy}}) \tag{58}$$

其中 $\sigma$ 表示激活函数. 则下一个词元的概率分布为

$$\alpha^{\text{next}}=p\times\alpha^{\text{copy}}+(1-p)\times\alpha^{\text{vocab}} \tag{59}$$

其中 $\alpha^{\text{copy}}$ 为实体和输入词元的分布，$\alpha^{\text{vocab}}$ 为词汇表上的分布. 在 BLEU 和 METEOR 的评估下，Graph-Writer 的性能明显优于 Rewriter 等方法.

长文本生成经过分层、规划、融合先验知识等方法的改善，已经可以应用于社交网络中一些简单的生成场景，但仍有许多需要进一步研究的问题. 比如当生成过程受控制时，生成文本的创造性和其内容的受控程度之间存在相互制约，且情节不连贯和逻辑不合理的问题仍没有得到解决. 此外，现有的一些质量评价方法度量能力较弱，不足以评价长文本的连贯性、逻辑性、创造性等指标. 因此，还需开发更统一和鲁棒的方法来衡量长文本的质量[95].

### 4.4 生成文本质量评价方法

社交网络中的文本要能够有效向阅读者传达所想表达的信息，还需要让阅读者有继续阅读的欲望，因此通常在可读性、逻辑性和主体一致性等方面都有较高的要求. 在这个前提下，选取适当的评价方法，对于生成的文本质量进行合理恰当的评价显得尤为重要. 社交网络的正常使用者均为人类，因此聘选人类阅读者对所生成的文本进行评价是一种自然的方式，但在应对大量的待评测文本时，人工评价时间成本高、经济成本高的劣势也尤为凸显. 为了应对更多文本质量评价的任务，人们对自动的、高效的、非人工的方法有较高的需求，因此适宜使用计算机评价生成文本质量的方法应运而生.

在生成文本质量评价中，一般的方法是将生成文本和参考文本进行对比，根据选定的评价方法特性使用合适的计算机制，最终根据评价方法对应的评价指标体系计算输出评价结果. 根据评价指标的数值，可以对生成文本质量进行客观系统性评价. 生成文本质量评价的基本流程如图 13 所示.

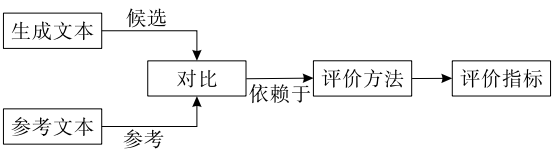


图 13 生成文本质量评价的基本流程

#### 4.4.1 基于 $n$ 元短语重叠的评价方法

2002 年 Papineni 等人[96] 提出的 BLEU，采用了一种根据数值度量来衡量机器(候选)翻译与人工(参考)翻译匹配程度的方法. 该方法通过统计候选翻译与参考翻译共同包含的 $n$ 元短语($n$ 个单词组成)在候选翻译中所占的比例(匹配精度)$p_n$，以及计算 $n$ 取不同值时匹配精度的加权 $w_n$ 几何平均值，可以很好地反映出句子的精确和通顺程度. 为了翻译长度的匹配，他们引入了一个简短惩罚因子

$$BP=\begin{cases}1 & c>r\\ e^{1-r/c}, & c\le r\end{cases} \tag{60}$$

其中，$c$ 表示候选翻译的文本长度，$r$ 示参考翻译的文本长度. 将匹配精度的几何平均值指数化后乘以惩罚因子，即可得到候选翻译的评分

$$\text{BLEU}=BP\exp\left(\sum_{n=1}^{N}w_n\log p_n\right) \tag{61}$$

BLEU 方法的评分与人工评分呈现出很好的线性关系，为后续的研究提供了有效的建模理念.

2004 年 Lin[73] 提出的 ROUGE 是一套用于评估长文本自动生成摘要的方法. 该方法主要考虑了四种主要的 $n$ 元短语计数方法(ROUGE-N、ROUGE-L、ROUGE-W 和 ROUGE-S)，用于测量生成文本与参考(人工书写)文本之间的 $n$ 元短语重叠. 其中，ROUGE-N 通过统计生成文本与单一参考文本之间的 $n$ 元短语重叠比例评价生成文本的质量. 当参考文本为多个时，ROUGE-N 返回最高的重叠比例作为对生成文本的评价. ROUGE-L 使用最长公共子序列评价文本的匹配程度. 由于该评价方法会自动寻找生成文本和参考文本匹配的最长 $n$ 元短语，因此不需要额外对 $n$ 元短语的长度进行预先定义. ROUGE-W 是在 ROUGE-L 的基础上对最长公共子序列的匹配情况进行了加权平均，解决了 ROUGE-L 无法对重叠字段语序相同而其余字段语序不同的生成文本进行有效区分的问题. ROUGE-S 使用生成文本和参考文本中任意间隔的二元短语重叠程度评价生成文本的质量，并且在此基础上提出了一个使用单个短语进行评价的方法 ROUGE-SU. 在 DUC 数据集上，ROUGE 评价方法与人工评分表现出了良好的相关性，并且随着参考文本数量的增加，这一方法与人工评分的相关程度也有所增加.

Banerjee 等人[97] 提出了一种名为 METEOR 的生成文本评价方法. 这是一种根据广义的单一短语匹配衡量机器翻译与人工(参考)翻译匹配程度的方法. 它包含三个顺序连接的匹配阶段，分别采用精确匹配、词干匹配和同义词匹配. 每一阶段仅对上一阶段后未完成匹配的单一短语进行匹配. 与 BLEU 方法不同，这一方法分别统计了匹配精度和匹配召回，并将两个值取调和平均值作为机器翻译和人工翻译匹配程度的 METEOR 评分，即

$$F_{\text{mean}}=\frac{10PR}{R+9P} \tag{62}$$

其中，$P$ 代表匹配精度，$R$ 代表匹配召回.

当处理长机器翻译文本的评分任务时，这一方法会先将长文本分为尽量少的文本块，使每一个文本块中每个单一短语均位置相邻. 它们映射到参考翻译后，参考翻译中的被映射短语也位置相邻，得分的惩罚因子定义如下：

$$Penalty=0.5\left(\frac{\text{文本块数}}{\text{完成匹配的单一短语数}}\right)^3 \tag{63}$$

最终的 METEOR 得分如下：

$$\text{Score}=F_{\text{mean}}\times(1-Penalty) \tag{64}$$

存在多个参考翻译时，METEOR 方法会使用所有参考翻译中最高的 METEOR 评分作为机器翻译的得分. 这一方法在 LDC TIDES 2003 数据集的阿拉伯语-英语和汉语-英语的翻译数据上进行了测试，取得了比 BLEU 和 NIST 更好的人工评分相关性. 一方面证实了匹配召回率比匹配精度具有更好的人工评分相关性，另一方面也证明了采用多种匹配方法可以对相关性取得有效的提升.

2020 年 Google 公司提出了基于 BERT 预训练模型的文本生成评价指标 BERTScore[98]. 他们通过计算两个句子嵌入词元的余弦相似度之和推断出两个句子之间的相似度. 该方法在保证语意相近的同时考虑了同义词和词组的多样性，并有效地保留了文本中相距较远的语意上的关系和顺序.

#### 4.4.2　基于距离的生成文本评价方法

字错误率[99]（Word Error Rate，WER）是一种基于编辑距离的评价方法，最初被设计用于测量语音识别系统的性能，也可以用于生成文本的质量评估. 它统计由翻译文本向参考文本转变所需要进行操作（包括替换、插入、删除）的总字数占参考文本的长度比例，

$$\text{WER}=\frac{\text{替换字数}+\text{插入字数}+\text{删除字数}}{\text{参考文本总长度}} \tag{65}$$

将这一比例值作为翻译文本到参考文本的距离. 字错误率评价方法的主要缺点在于它的评价结果依赖参考文本，并且当存在多个不同的参考文本时，这一方法只能选择一个参考文本作为基准.

字错误率的一种变体评价方法是多参考字错误率[100]（Multi-reference Word Error Rate，m-WER）评价，这一评价方法于 2016 年由 Ali 等人提出. 这种评价方法针对同一生成文本可以在多个参考文本上分别进行字错误率评价，得到多个不同的距离结果后，选择距离结果的最小值作为生成文本到参考文本的距离. 这一方法的主要缺陷在于需要人工获取多个参考文本.

翻译编辑率[76]（Translation Edit Rate，TER）是一种面向多个参考文本的评价方法，由 Snover 等人于 2006 年提出. 它被定义为由翻译文本通过编辑向多个参考文本中的一个转变时所需要进行的编辑次数，除以多个参考文本的平均字数. 根据编辑的最小次数，TER 将取为翻译文本到最接近的参考文本的翻译操作率，

$$\text{TER}=\frac{\text{编辑次数}}{\text{多个参考文本的平均字数}} \tag{66}$$

这一方法考虑了单个字的替换、插入、删除以及字词形式的转换，并且考虑了多个连续字词在文本中移动的情况. 与人工判断的结果相比，这一方法取得了良好的关联性，但仍存在一些局限. 例如，在处理文本时，它更多考虑和参考文本间精确的字词匹配，对于同义词等情况不能做出合理的评价. 在此基础上，一些 TER 评价方法的变体做出了改进，如 TERP① 在计算编辑次数时引入了短语替换、词干提取等方法，并对字词的位置调整约束进行了放宽；ITER[101] 则额外增加了词干匹配和标准化.

## 5　数据集

本节对目前开源的相关数据集进行了总结，为在线社交网络对抗技术中的机器文本检测与生成方法提供有效的训练与验证数据集. 各数据集特征信息与应用场景如表 4 所示.

**CBT（Children's Book Test，儿童书籍测试）**[102]. 这是 2015 年公布的一个儿童书本内容数据集. 它由来自古腾堡计划（Project Gutenberg）的 108 本免费儿童书籍构成，使用书本段落中的连续语句组成了文本材料，并在材料后附带了问题和备选答案. 这一数据集包含了近 70 万个不同的问题，所使用词汇表包含 53 628 个词. 设计这一数据集的主要目的是衡量语言模型根据上下文信息回答相关问题的能力. 使用儿童书籍作为文本材料来源，确保了材料有清晰的叙事结构，从而使上下文信息的作用更加突出.

**Books**[103]. 这是 2015 年公布的一个由书籍组成的语料库，包含了 11 038 本未公开作者的免费书籍，每本书籍字数超过 2 万字. 这一数据集包含了

① TERP 方法以马里兰大学的吉祥物 Terrapin 命名.

**表 4 数据集特征总结**

| 应用场景 | 数据集名称 | 数据规模 |
|---|---|---|
| 长文本，应用于社交网络中新闻、网页文章等场景 | CBT(Children's Book Test，儿童书籍测试)[102] | 近 70 万个基于文本材料的问题和相应的备选答案 |
| | Books[103] | 11038 本未公开作者的书籍，每本书超过 2 万字，涵盖 16 种不同类型 |
| | Wikitext[104] | 英文维基百科上 23805 篇“好文章”和 4790 篇“特色文章” |
| | 莎士比亚剧集[65] | 39 部戏剧、154 首十四行诗以及 2 首长叙事诗 |
| | RealNews[18] | 从 Commen Crawl 中收集的新闻文章 |
| | 奥巴马演讲 | 11092 个段落，73 万余个单词 |
| | ROCstories[87] | 49255 篇高质量短篇故事 |
| | Politic-News | 关于政治新闻的真实文本 |
| | WebText[4] | 来自 450 万个网络链接的超过 800 万份文本 |
| 短文本，应用于社交网络中短评、评价等场景 | 中文诗集[63] | 28.5 万篇诗词，超过 272 万行诗句 |
| | IMDB 文本语料库[105] | 35 万条影评 |
| | SST-full[106] | 从影评中提取的 21.5 万个独立短语 |
| | 中文诗集 2 | 16394 首中文五言绝句 |
| 语义理解，应用于社交网络中总结、转述及翻译等场景 | DUC[107] | 会议原始文本和会议摘要 |
| | LDC TIDES 2003[108] | 不同语种向英语翻译的原始文本和翻译参考文本 |
| | E2E[83] | 对应于 6039 种含义表达的 51426 个人类语言样本 |
| 交互文本，应用于社交网络中问答、人机交互等场景 | Quora | 超过 40 万行问题对(问题 1，问题 2)，并对每对问题是否属于同一个问题进行了标记 |
| | PersonageNLG[82] | 使用了 PERSONAGE 生成的对话样本，有 88855 个训练样本和 1390 个测试样本 |
| 图片标注，应用于社交网络中图文组合场景 | MSCOCO[64] | 91 种常见物体的图片，其中有 82 种有超过 5000 个经过人工标注的带标签样本 |

16 种不同类型的书籍，总句数超过 7400 万句，包含近 10 亿单词. GPT-3 模型将这一数据集作为高质量参考语料库加入到了训练集组合中，用以增强模型生成文本时的多样性.

**Wikitext**[104]. 将英文维基百科上 23805 篇“好文章”和 4790 篇“特色文章”作为文本材料. 这些文章经过人工审查，被认为写作良好、事实准确、覆盖广泛、观点中立. 在组建词汇表时，出现次数低于 3 次的单词被抛弃，并统一映射到了〈unk〉标记. 整个数据集包含 1.03 亿单词，并提供了 WikeText-2 和 WikiText-103 两种版本：WikeText-2 包含较少的训练文本，词汇表也相对较小；WikiText-103 包含了全部的提取文本，也拥有完整的词汇表.

**中文诗集**[63]. 这是一个根据在线资源搜集整理的中文古典诗数据集，包含唐代以来的中文古典诗词. 数据集中总计有 28.5 万篇诗词，其中绝句有 78859 篇. 整个数据集包含超过 272 万行诗句，总计约 1572 万余字.

**莎士比亚剧集**[65]. 包含了莎士比亚的所有作品，具体包括 39 部戏剧、154 首十四行诗以及 2 首长叙事诗，其中含有莎士比亚自创词.

**RealNews**[18]. 从 Commen Crawl 中收集的新闻文章，包含 120 GB 的语料库. Grover 模型将这一数据集根据文章发布时间划分为了训练集和验证集，使用 2016 年 12 月至 2019 年 3 月的文章数据作为训练样本，使用 2019 年 4 月发表的文章作为验证样本. 在训练过程中，Grover 采用了随机采样的方法，从文章数据中随机选取长度为 1024 的序列进行训练.

**IMDB 文本语料库**[105]. 这个数据集包含了从 IMDb 网站提取的来自 5 万部电影的影评信息. 在所有影评信息中，累计评论量少于两个的用户的有关评论被移除，总评论量少于两个的电影的用户有关条目也被移除. 经过筛选后，IMDb 数据集包含了针对 22380 部电影，来自 54671 名用户的近 35 万条观影评论.

**SST-full**[106]. 这是第一个带有完整标记的解析树语料库. 语料库包含从电影评论中提取的 10662 个单句，使用斯坦福解析器将所有单句分割成了 21.5 万个独立短语. 通过在 AMT(Amazon Mechanical Turk)论坛上发布标注任务，每个短语由三名人类标注者进行了人工标注.

**中文诗集 2**(见本文 P12 脚注①). 包含 16394 首汉语五言绝句，每首诗含有 20 个汉字，用于 SeqGAN 的训练.

**奥巴马演讲**(见本文 P12 脚注②). 选自奥巴马所有演讲内容，包含 11092 个段落，有 73 万余个单词，用于 SeqGAN 的训练.

**ROCstories**[87]. 这一数据集包含了 49255 篇高质量短篇故事. 每篇故事由长度不超过 70 个词的句子组成，包含日常事件间各种常识性的因果关系和时间关系，并由三个人审核确保了故事具有较高的

质量和足够的真实性.故事搜集通过在 AMT 论坛上发布众包任务完成,平均每名工作者提供了 52 篇故事.

**DUC**[107]. 包含会议原始文本和会议摘要,根据摘要生成方式可分为人工编写的会议概要、自动生成的基准性总结文本、志愿小组提交的摘要文本等.在摘要文本长度方面,包含长度约 10 个词的单文本摘要,也包含长度约 100 词的多文本摘要.要使用这一数据集,需要填写申请表单向 NIST(National Institute of Standards and Technology,美国国家标准与技术研究院)发出申请,获得许可后可以下载使用.

**LDC TIDES 2003**[108]. 这一数据集包含由不同语种向英语翻译的原始文本和翻译参考文本.其中的中文数据集包含 920 句中文语句,阿拉伯语数据集包含 664 句阿拉伯语文本,每一条原始语句提供了四条对应的翻译参考语句.

**WebText**[4]. 包含来自 450 万个网络链接的文本信息,总计超过 800 万份文本数据,用来对 GPT-2 模型进行参数训练,GPT-3 模型也在这一数据集上进行了训练.

**MSCOCO**[64]. 包含 91 个常见物体类别,其中有 82 个类别有超过 5000 个经过人工标注的带有标签的样本.图像搜集于日常场景中,由生活中的常见物体和自然背景组成,可以被 4 岁儿童轻易识别.数据集被划分为训练集、验证集和测试集三个部分,训练集包含 16.5 万张图片,验证集包含 81 208 张图片,测试集包含 81 434 张图片.

**Quora**①. 由超过 40 万条问题对(问题 1,问题 2)组成,并对这两个问题是否是同一个问题进行了标记.

**Politic-News**②. Uchendu 等人[41]提供的关于政治新闻的真实文本,并包含分别由 CTRL、GPT-1、GPT-2、GROVER、XLM[109]、XLNet、PPLM 和 FAIR[110]等模型生成的文本.

**E2E**[83]. 包含 51 426 个人类语言样本,对应于 6039 种含义表达.其中训练集包含 42 061 个样本,含义表达为 4862 种.开发集包含对应于 547 个含义表达的 4672 种样本.测试集具有 4693 个样本,对应 630 种含义表达.

**PersonageNLG**[82]. 包含使用了 PERSONAGE 生成的对话样本,有 88 855 个训练样本和 1390 个测试样本.PERSONAGE 使用材料的含义表达和参数文件控制对话生成,在这一数据集中,使用了 8 种不同的含义表达和 5 种风格属性生成对话.

## 6　未来研究方向

基于本文对社交网络文本内容的自动检测与生成工作总结,作者对当前工作的不足和该方向未来重要的研究方向进行了总结.

### 6.1　增强检测与生成方法的泛化能力

社交网络文本种类多样且内容丰富,这使得社交网络文本内容的检测和生成模型需要具备较强的泛化能力,以应用于丰富的在线社交网络对抗应用场景.但是现有的文本检测与生成方法对模型的设定与训练数据集有非常强的依赖性.例如,现有生成文本检测方法往往只能检测采用特定设置的特定模型生成的文本.当模型结构、采样方法、模型大小、训练数据等设置发生变化时,已有的检测器与生成器很难保持原有较好的算法性能[111].另外,现有的方法在内容领域方面的泛化性较差,Bakhtin 等人发现现有检测方法无法应用于多个不同平台(如维基百科、书籍、新闻)的文本[56].因此,增强机器生成文本检测方法与文本自动生成方法的泛化能力变得势在必行.

### 6.2　增强检测与生成方法的可解释性

在检测和生成长度较短、语义信息不够丰富的社交网络文本时,深度学习模型需要与人类协作才可以达到更好的效果.这需要机器模型具有一定的可解释性,为人类与机器协作提供参考.但是深度学习方法的可解释性一直是大家非常关心的开放性问题.受 GLTR 的启发,无论是检测器还是生成器,都应提供更多的可解释信息,对其输出结果进行解释.这样有助于人和机器协同配合完成检测与生成任务,提高检测与生成文本的准确度.根据方法的可解释性,人们可以给出接受或反对检测与生成结果的理由,并通过反馈降低检测与生成不准确可能带来的危害.

### 6.3　基于外部知识图谱的事实核查机制

在线社交网络对抗场景中的特定文本生成往往要求包含大量的事实信息,仅仅依靠深度学习模型挖掘文本中的概率分布差异会导致忽略文本信息的真实性,也使得检测模型不够"智能".例如,Schuster 等

① https://www.quora.com/q/quoradata/First-Quora-Dataset-Release-Question-Pairs

② https://github.com/AdaUchendu/Authorship-Attribution-for-Neural-Text-Generation

人[39]的研究表明，现有的检测器在检测机器生成的真实文本和人为改动真实文本产生的虚假文本时，判断准确率并不高. 其原因在于检测器过于依赖文本的分布特征，而这些分布特征并不能用来区分来源相似的文本. 目前，基于外部知识图谱的先验知识在信息检索、推荐系统、身份对齐等领域开始逐步被使用，且取得了很好的效果. 因此，在社交网络文本生成与检测过程中，可以将模型与已有的知识图谱结合起来，建立借助知识库评估文本真实性的方法，即事实核查机制，达到文本的精准生成与检测.

### 6.4 增强检测与生成方法的鲁棒性

社交网络文本存在不规范性，例如微博、Twitter等社交网络平台上发布的文本，常常存在表情符号多、有错别字、或者新兴流行语的情况. Wolff[112]的研究发现，使用简单的策略，如替换同型词或者错误拼写一些单词，可以对 RoBERTa 检测器进行有效的攻击，使其检测结果出错. 这反映了现有检测器较为脆弱，容易在轻微扰动下做出不同的判断. 同时，目前已有的文本生成模型大都依赖于基于特定数据集的精调参数，轻微的设置变化对算法性能都会造成比较严重的影响. 例如 Wallace 等人[113]对“通用攻击”方法进行研究，试图通过对抗样本来找出检测器的缺陷，以便设计出鲁棒性更强的检测器来对抗不同种类的攻击. 因此，亟需增强社交网络文本检测与自动生成方法的鲁棒性，以提高当前严峻网络空间多变环境下防御的稳定性.

### 6.5 增强语言模型的易用性

内容产生与内容接收的广泛性也是在线社交网络文本的一大重要特性，这意味着人们需要能够较为轻松的创作内容与接收内容. 目前，虽然预训练语言模型的出现，大大降低了社交网络中机器文本的检测与自动生成的门槛. 但是使用这些现有的语言模型需要具备一定的机器学习知识基础与编程技巧，并需要具备可用的训练与测试数据集. 开发类似于 Talk to Transformer 和 Write with Transformer 等与语言模型交互的工具可以增强语言模型的易用性，扩大语言模型的应用范围和用户规模. 同时，语言模型的大规模应用与用户反馈有助于提升模型的表现，在广大用户群体使用过程中也会产生更多有创意的想法.

## 7 总 结

本篇综述对社交网络对抗中的两个重要研究方向——机器生成文本检测和文本自动生成最新的研究成果进行了详尽的调研. 我们对文本检测方法结合在线社交网络的特点，从零次分类模型、预训练模型再训练、检测机器特征、能量基础模型和人机协作模型等几个方向进行了详细分析. 对文本生成方式从生成对抗网络、可控文本生成、长文本生成、文本质量的评价方式四个角度进行了总结和优缺点分析. 关于可控文本生成，将其方法分为控制标记和属性分类器两类. 本文对相关工作的综述可以帮助读者快速了解在线社交网络对抗技术的国内外研究现状，可以帮助读者在大量的相关研究文献中快速理解问题背景. 同时，我们也提出了多个重要的未来研究方向，希望可以对社交网络对抗领域未来的研究工作提供指导和帮助. 希望我们的工作可以阻止恶意用户滥用语言模型，达到维护网络空间内容安全的目的.

## 参 考 文 献


[1] Liu X, Shen C, Guan X, et al. DIGGER: Detect similar groups in heterogeneous social networks. ACM Transactions on Knowledge Discovery from Data, 2018, 13(1): 1-27

[2] Ferrara E, Varol O, Davis C, et al. The rise of social bots. Communications of the ACM, 2016, 59(7): 96-104

[3] Devlin J, Chang M, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019). Minneapolis, USA, 2019: 4171-4186

[4] Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners. OpenAI Blog, 2019, 1(8): 9

[5] Brown T B, Mann B, Ryder N, et al. Language models are few-shot learners. arXiv preprint arXiv:2005.14165, 2020

[6] Fan A, Lewis M, Dauphin Y. Hierarchical neural story generation//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018). Melbourne, Australia, 2018: 889-898

[7] Majumder B P, Li Shuyang, Ni Jianmo, McAuley J. Interview: Large-scale modeling of media dialog with discourse patterns and knowledge grounding//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020: 8129-8141

[8] Wang W, Li P, Zheng H T. Consistency and coherency enhanced story generation. arXiv preprint arXiv:2010.08822, 2020

[9] Zhang Y, Sun S, Galley M, et al. DialoGPT: Large-scale generative pre-training for conversational response generation. arXiv preprint arXiv:1911.00536, 2019

[10] Gupta P, Bigham J P, Tsvetkov Y, et al. Controlling dialogue generation with semantic exemplars. arXiv preprint arXiv:2008.09075, 2020

[11] Wu S, Li Y, Zhang D, et al. Diverse and informative dialogue generation with context-specific commonsense knowledge awareness//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online, 2020: 5811-5820

[12] Wan X, Zhang J. CTSUM: Extracting more certain summaries for news articles//Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'14). Gold Coast, Australia, 2014: 787-796

[13] Zhang J, Zhao Y, Saleh M, et al. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization//Proceedings of the 37th International Conference on Machine Learning. Virtual Event, 2020: 11328-11339

[14] Zhou Jian, Tian Xuan, Cui Xiao-Hui. Generation method of text summarization based on advanced Sequence-to-Sequence model. Computer Engineering and Applications, 2019, 55(1): 128-134(in Chinese)
(周健，田萱，崔晓晖. 基于改进 Sequence-to-Sequence 模型的文本摘要生成方法. 计算机工程与应用，2019，55(1)：128-134)

[15] Liu G, Hsu T M H, Mcdermott M, et al. Clinically accurate chest X-ray report generation//Proceedings of the Machine Learning for Healthcare Conference. Ann Arbor, USA, 2019: 249-269

[16] Biswal S, Xiao C, Glass L, et al. Clinical report auto-completion//Proceedings of the Web Conference 2020. Taipei, China, 2020: 541-550

[17] Feng Xiao-Cheng, Gong Heng, Leng Hai-Tao, et al. Extractive essay generation for college entrance examination. Chinese Journal of Computers, 2020, 43(2): 315-325 (in Chinese)
(冯骁骋，龚恒，冷海涛等. 基于抽取的高考作文生成. 计算机学报，2020，43(2)：315-325)

[18] Zellers R, Holtzman A, Rashkin H, et al. Defending against neural fake news//Advances in Neural Information Processing Systems. Vancouver, Canada, 2019: 9054-9065

[19] Yanagi Y, Orihara R, Sei Y, et al. Fake news detection with generated comments for news articles//Proceedings of the 2020 IEEE 24th International Conference on Intelligent Engineering Systems (INES). 2020: 85-90

[20] Adelani D I, Mai H, Fang F, et al. Generating sentiment-preserving fake online reviews using neural language models and their human- and machine-based detection//Proceedings of the International Conference on Advanced Information Networking and Applications. Caserta, Italy, 2020: 1341-1354

[21] Tan E, Guo L, Chen S, et al. Spammer behavior analysis and detection in user generated content on social networks//Proceedings of the 2012 IEEE 32nd International Conference on Distributed Computing Systems. Macau, China, 2012: 305-314

[22] Grinberg N, Joseph K, Friedland L, et al. Fake news on twitter during the 2016 us presidential election. Science, 2019, 363(6425): 374-378

[23] Adams T. AI-powered social bots. arXiv preprint arXiv: 1706.05143, 2017

[24] Chen Yan-Fang, Li Zhi-Yu, Liang Xun, et al. Review on rumor detection of online social networks. Chinese Journal of Computers, 2018, 41(7): 1648-1677(in Chinese)
(陈燕方，李志宇，梁循等. 在线社会网络谣言检测综述. 计算机学报，2018，41(7)：1648-1677)

[25] Elman J L. Finding structure in time. Cognitive Science, 1990, 14(2): 179-211

[26] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8): 1735-1780

[27] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need//Advances in Neural Information Processing Systems. Long Beach, USA, 2017: 5998-6008

[28] Pan S J, Yang Q. A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering, 2009, 22(10): 1345-1359

[29] Xu Bing-Bing, Cen Ke-Ting, Huang Jun-Jie, et al. A survey on graph convolutional neural network. Chinese Journal of Computers, 2020, 43(5): 755-780(in Chinese)
(徐冰冰，岑科廷，黄俊杰等. 图卷积神经网络综述. 计算机学报，2020，43(5)：755-780)

[30] Bruna J, Zaremba W, Szlam A, et al. Spectral networks and locally connected networks on graphs. arXiv preprint arXiv: 1312.6203, 2013

[31] Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv: 1609.02907, 2016

[32] Veličković P, Cucurull G, Casanova A, et al. Graph attention networks. arXiv preprint arXiv:1710.10903, 2017

[33] Kingma D P, Welling M. Auto-encoding variational Bayes. Stat, 2014, 1050: 1

[34] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets//Advances in Neural Information Processing Systems. Montreal, Canada, 2014: 2672-2680

[35] Gehrmann S, StrobelT H, Rush A M. GLTR: Statistical detection and visualization of generated text//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Florence, Italy, 2019: 111-116

[36] Hashimoto T, Zhang H, Liang P. Unifying human and statistical evaluation for natural language generation//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, USA, 2019: 1689-1701

[37] Holtzman A, Buys J, Du L, et al. The curious case of neural text degeneration. arXiv preprint arXiv:1904.09751, 2019

[38] Ippolito D, Duckworth D, Callison-Burch C, et al. Automatic detection of generated text is easiest when humans are fooled //Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online, 2020: 1808-1822

[39] Schuster T, Schuster R, Shah D J, et al. The limitations of stylometry for detecting machine-generated fake news. Computational Linguistics, 2020, 46(2): 499-510

[40] Joulin A, Grave É, Bojanowski P, et al. Bag of tricks for efficient text classification//Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. Valencia, Spain, 2017: 427-431

[41] Uchendu A, Le T, Shu K, et al. Authorship attribution for neural text generation//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online, 2020: 8384-8395

[42] Solaiman I, Brundage M, Clark J, et al. Release strategies and the social impacts of language models. arXiv preprint arXiv:1908.09203, 2019

[43] Fagni T, Falchi F, Gambini M, et al. TweepFake: About detecting deepfake tweets. arXiv preprint arXiv: 2008.00036, 2020

[44] Keskar N S, Mccann B, Varshney L R, et al. CTRL: A conditional transformer language model for controllable generation. arXiv preprint arXiv:1909.05858, 2019

[45] Shu L, Papangelis A, Wang Y C, et al. Controllable text generation with focused variation//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings. Online Event, 2020: 3805-3817

[46] Tay Y, Bahri D, Zheng C, et al. Reverse engineering configurations of neural text generation models//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online, 2020: 275-279

[47] Zhong W, Tang D, Xu Z, et al. Neural deepfake detection with factual structure of text//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. Online, 2020: 2461-2470

[48] Liu Y, Ott M, Goyal N, et al. RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv: 1907.11692, 2019

[49] Yamada I, Asai A, Sakuma J, et al. Wikipedia2Vec: An efficient toolkit for learning and visualizing the embeddings of words and entities from Wikipedia//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Online, 2020: 23-30

[50] Yang Z, Dai Z, Yang Y, et al. XLNet: Generalized autoregressive pretraining for language understanding//Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019. Vancouver, Canada, 2019: 5754-5764

[51] Jawahar G. Detecting human written text from machine generated text by modeling discourse coherence. https://www.cs.ubc.ca/~carenini/TEACHING/CPSC503-20/SOME-FINAL-PROJECTS/GANESH-503_Report-GC.pdf

[52] Lai A, Tetreault J. Discourse coherence in the wild: A dataset, evaluation and methods//Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue. Melbourne, Australia, 2018: 214-223

[53] Barzilay R, Lapata M. Modeling local coherence: An entity-based approach. Computational Linguistics, 2008, 34(1): 1-34

[54] Guinaudeau C, Strube M. Graph-based local coherence modeling//Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. Sofia, Bulgaria, 2013: 93-103

[55] LeCun Y, Chopra S, Hadsell R, et al. A Tutorial on Energy-Based Learning. Predicting Structured Data, Cambridge: MIT Press, 2006

[56] Bakhtin A, Gross S, Ott M, et al. Real or fake? Learning to discriminate machine from human generated text. arXiv preprint arXiv:1906.03351, 2019

[57] Dauphin Y N, Fan A, Auli M, et al. Language modeling with gated convolutional networks//Proceedings of the 34th International Conference on Machine Learning. Sydney, Australia, 2017: 933-941

[58] Dugan L, Ippolito D, Kirubarajan A, et al. RoFT: A tool for evaluating human detection of machine-generated text//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Online, 2020: 189-196

[59] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks. Communications of the ACM, 2020, 63(11): 139-144

[60] Zhang Y, Gan Z, Carin L. Generating text via adversarial training//Proceedings of the NIPS Workshop on Adversarial Training. Barcelona, Spain, 2016: 21-32

[61] Yu L, Zhang W, Wang J, et al. SeqGAN: Sequence generative adversarial nets with policy gradient//Proceedings of the 31st AAAI Conference on Artificial Intelligence. San Francisco, USA, 2017: 2852-2858

[62] Lin K, Li D, He X, et al. Adversarial ranking for language generation. Advances in Neural Information Processing Systems, 2017, 30: 3155-3165

[63] Zhang X, Lapata M. Chinese poetry generation with recurrent neural networks//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha, Qatar, 2014: 670-680

[64] Lin T Y, Maire M, Belongie S, et al. Microsoft COCO: Common objects in context//Proceedings of the 13th European Conference on Computer Vision (ECCV 2014). Zurich, Switzerland, 2014: 740-755

[65] Shakespeare W. The Complete Works of William Shakespeare. Hertfordshire: Wordsworth Editions, 2007

[66] Guo J, Lu S, Cai H, et al. Long text generation via adversarial training with leaked information//Proceedings of the

32nd AAAI Conference on Artificial Intelligence(AAAI-18), the 30th Innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence(EAAI-18). New Orleans, USA, 2018: 5141-5148

[67] Vezhnevets A S, Osindero S, Schaul T, et al. Feudal networks for hierarchical reinforcement learning//Proceedings of the 34th International Conference on Machine Learning. Sydney, Australia, 2017: 3540-3549

[68] Dathathri S, Madotto A, Lan J, et al. Plug and play language models: A simple approach to controlled text generation//Proceedings of the 8th International Conference on Learning Representations. Addis Ababa, Ethiopia, 2020

[69] See A, Roller S, Kiela D, et al. What makes a good conversation? How controllable attributes affect human judgments//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, USA, 2019: 1702-1723

[70] Chan A, Ong Y S, Pung B, et al. CoCon: A self-supervised approach for controlled text generation//Proceedings of the 9th International Conference on Learning Representations. Virtual Event, Austria, 2021

[71] Lin C Y, Och F J. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics//Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics. Barcelona, Spain, 2004: 605-612

[72] Lavie A, Agarwal A. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments//Proceedings of the 2nd Workshop on Statistical Machine Translation. Prague, Czech Republic, 2007: 228-231

[73] Lin C Y. ROUGE: A package for automatic evaluation of summaries//Proceedings of the Text Summarization Branches Out. Barcelona, Spain, 2004: 74-81

[74] Sohn K, Lee H, Yan X. Learning structured output representation using deep conditional generative models//Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015. Montreal, Canada, 2015: 3483-3491

[75] Gupta A, Agarwal A, Singh P, et al. A deep generative framework for paraphrase generation//Proceedings of the 32nd AAAI Conference on Artificial Intelligence(AAAI-18), the 30th Innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence(EAAI-18). New Orleans, USA, 2018: 5149-5156

[76] Snover M, Dorr B, Schwartz R, et al. A study of translation edit rate with targeted human annotation//Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers. Cambridge, USA, 2006: 223-231

[77] Nair V, Hinton G E. Rectified linear units improve restricted Boltzmann machines//Proceedings of the 27th International Conference on Machine Learning. Haifa, Israel, 2010: 807-814

[78] Ba J L, Kiros J R, Hinton G E. Layer normalization. Stat, 2016, 1050: 21

[79] Child R, Gray S, Radford A, et al. Generating long sequences with sparse transformers. arXiv preprint arXiv: 1904.10509, 2019

[80] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 770-778

[81] Van Den Oord A, Vinyals O, et al. Neural discrete representation learning//Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017. Long Beach, USA, 2017: 6306-6315

[82] Oraby S, Reed L, Tandon S, et al. Controlling personality-based stylistic variation with neural natural language generators//Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue. Melbourne, Australia, 2018: 180-190

[83] Dušek O, Novikova J, Rieser V. Evaluating the state-of-the-art of end-to-end natural language generation: The E2E NLG challenge. Computer Speech & Language, 2020, 59: 123-156

[84] Peng N, Ghazvininejad M, May J, et al. Towards controllable story generation//Proceedings of the First Workshop on Storytelling. New Orleans, USA, 2018: 43-49

[85] Tang D, Qin B, Liu T. Document modeling with gated recurrent neural network for sentiment classification//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal, 2015: 1422-1432

[86] Rose S, Engel D, Cramer N, et al. Automatic keyword extraction from individual documents. Text Mining: Applications and Theory, 2010, 1: 1-20

[87] Mostafazadeh N, Chambers N, He X, et al. A corpus and cloze evaluation for deeper understanding of commonsense stories//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, USA, 2016: 839-849

[88] Hu Z, Yang Z, Liang X, et al. Toward controlled generation of text//Proceedings of the 34th International Conference on Machine Learning. Sydney, Australia, 2017: 1587-1596

[89] Hinton G E, Dayan P, Frey B J, et al. The "wake-sleep" algorithm for unsupervised neural networks. Science, 1995, 268(5214): 1158-1161

[90] Nguyen A, Clune J, Bengio Y, et al. Plug & play generative networks: Conditional iterative generation of images in latent space//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 3510-3520

[91] Yao L, Peng N, Weischedel R, et al. Plan-and-Write: Towards better automatic storytelling//Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence, the Thirty-First Innovative Applications of Artificial Intelligence Conference, the Ninth AAAI Symposium on Educational Advances in Artificial Intelligence. Honolulu, USA, 2019: 7378-7385

[92] Xu P, Patwary M, Shoeybi M, et al. Controllable story generation with external knowledge using large-scale language models//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online, 2020: 2831-2845

[93] Cer D, Yang Y, Kong S Y, et al. Universal sentence encoder for English//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018): System Demonstrations. Brussels, Belgium, 2018: 169-174

[94] Koncel-Kedziorski R, Bekal D, Luan Y, et al. Text generation from knowledge graphs with graph transformers//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, USA, 2019: 2284-2293

[95] Alabdulkarim A, Li S, Peng X. Automatic story generation: Challenges and attempts. arXiv preprint arXiv:2102.12634, 2021

[96] Papineni K, Roukos S, Ward T, et al. BLEU: A method for automatic evaluation of machine translation//Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Philadelphia, USA, 2002: 311-318

[97] Banerjee S, Lavie A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments//Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization @ ACL 2005. Ann Arbor, USA, 2005: 65-72

[98] Zhang T, Kishore V, Wu F, et al. BERTScore: Evaluating text generation with BERT//Proceedings of the 8th International Conference on Learning Representations. Addis Ababa, Ethiopia, 2020

[99] Tomás J, Mas J À, Casacuberta F. A quantitative method for machine translation evaluation//Proceedings of the EACL 2003 Workshop on Evaluation Initiatives in Natural Language Processing: Are Evaluation Methods, Metrics and Resources Reusable? Columbus, USA, 2003: 27-34

[100] Ali A, Magdy W, Bell P, et al. Multi-reference WER for evaluating ASR for languages with no orthographic rules//Proceedings of the 2015 IEEE Workshop on Automatic Speech Recognition and Understanding. Scottsdale, USA, 2015: 576-580

[101] Panja J, Naskar S K. ITER: Improving translation edit rate through optimizable edit costs//Proceedings of the Third Conference on Machine Translation: Shared Task Papers. Belgium, Brussels, 2018: 746-750

[102] Hill F, Bordes A, Chopra S, et al. The goldilocks principle: Reading children's books with explicit memory representations. arXiv preprint arXiv:1511.02301, 2015

[103] Zhu Y, Kiros R, Zemel R, et al. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books//Proceedings of the 2015 IEEE International Conference on Computer Vision. Santiago, Chile, 2015: 19-27

[104] Merity S, Xiong C, Bradbury J, et al. Pointer sentinel mixture models. arXiv preprint arXiv:1609.07843, 2016

[105] Diao Q, Qiu M, Wu C Y, et al. Jointly modeling aspects, ratings and sentiments for movie recommendation (JMARS)//Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA, 2014: 193-202

[106] Socher R, Perelygin A, Wu J, et al. Recursive deep models for semantic compositionality over a sentiment treebank//Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP, Grand Hyatt Seattle. Seattle, USA, 2013: 1631-1642

[107] Over P. An introduction to DUC 2003: Intrinsic evaluation of generic news text summarization systems//Proceedings of the Document Understanding Conference 2003. https://paperswithcode.com/dataset/duc-2004, 2003

[108] Ferro L, Gerber L, Mani I, et al. TIDES: 2003 standard for the annotation of temporal expressions. Mitre Corp Mclean Va Mclean, 2003

[109] Conneau A, Lample G. Cross-lingual language model pretraining//Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019. Vancouver, Canada, 2019: 7059-7069

[110] Chen P J, Shen J, Le M, et al. Facebook AI's WAT19 Myanmar-English translation task submission//Proceedings of the 6th Workshop on Asian Translation. Hong Kong, China, 2019: 112-122

[111] Bakhtin A, Deng Y, Gross S, et al. Energy-based models for text. arXiv preprint arXiv:2004.10188, 2020

[112] Wolff M. Attacking neural text detectors. arXiv preprint arXiv:2002.11768, 2020

[113] Wallace E, Feng S, Kandpal N, et al. Universal adversarial triggers for attacking and analyzing NLP//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong, China, 2019: 2153-2162

**LIU Xiao-Ming**, Ph. D., associate professor. His research interests include big graph mining, large-scale heterogeneous data analysis and mining, online social network adversarial technology, machine learning and its applications.

**ZHANG Zhao-Han**, M. S. candidate. His research interests include online social network adversarial technology and machine-generated text detection.

**YANG Chen-Yang**, M. S. candidate. His research interests include online social network adversarial technology and controllable text auto-generation.

**ZHANG Yu-Chen**, M. S. candidate. His research interests include online social network adversarial technology and machine-generated text detection.

**SHEN Chao**, Ph. D., professor. His research interests include cyber-physical system optimization and security, network and system security, and artificial intelligence security.

**ZHOU Ya-Dong**, Ph. D., associate professor. His research interests include data analysis and mining, network science and its applications.

**GUAN Xiao-Hong**, Ph. D., professor, Academician of Chinese Academy of Sciences. His research interests include allocation and scheduling of complex networked resources, network security, and sensor networks.

**Background**

With the development of the Internet, online social network adversarial technology has attracted attention from researchers in the field of artificial intelligence and cyberspace security. It is the foundation for protecting the users from the organized attackers and realizing countermeasures of social platform abnormal information delivery. Although online social network adversarial technology is a new concept with few related work, existing machine learning methods can be employed to solve practical problems in this new field. This paper mainly summarized the outstanding related work about the social network adversarial technology from two aspects, including the machine-generated text detection and content auto-generation. Besides, this paper thoroughly and systematically summarizes the relevant datasets to benefit the readers to verify the validity of the model. Finally, this paper summarizes the important research directions and challenges of online social network adversarial technology in the future. The detailed summary analysis of advantages and disadvantages for state-of-arts could provide a reference for future research of the readers.

This work was supported in part by the National Natural Science Foundation of China under Grant Nos. 61902308, 62103323, 61822309, 61773310, U1736205, U1766215, the Initiative Postdocs Supporting Program under Grant Nos. BX20190275, BX20200270, the China Postdoctoral Science Foundation under Grant Nos. 2019M663723, 2021M692565, and the Foundation of Xi'an Jiaotong University under Grant Nos. xjh032021058, xxj022019016, xtr022019002.